A Survey of Unified Multimodal Understanding and Generation: Advances and Challenges

Yan Yang[1,*], Haochen Tian[2,*,†], Yang Shi[3,*], Wulin Xie[2,*], Yi-Fan Zhang[2,‡]
Yuhao Dong[4], Yibo Hu[2], Liang Wang[2], Ran He[2], Caifeng Shan[1], Chaoyou Fu[1,‡], Tieniu Tan[1]

Abstract—Advancing AGI requires AI that can jointly understand and generate across modalities—text, images, video, and audio. As illustrated in Fig. 1, this unification evolves through three stages: from isolated expertise with separate models, to integrated capabilities in a unified framework, to emergent behaviors as a future vision enabling complex interleaved reasoning. This unification is motivated by two factors: (1) mutual reinforcement, where strong comprehension enables high-quality creation, and generation aids difficult reasoning via feedback loops; and (2) the flexibility to tackle complex real-world problems, such as turning a script into a coherent movie—something isolated models cannot handle. Despite promising open-source efforts (e.g., BAGEL, Emu3) and powerful closed-source models (e.g., GPT-4o, Gemini 2.0 Flash) demonstrating comparable performance, open-source unified foundation models (UFMs) still lag behind closed-source counterparts. The open-source community lacks consensus on key design choices for UFMs, such as modeling paradigms (autoregressive vs. hybrid diffusion), tokenizers, training strategies, and data curation, hampering progress. To bridge this gap, we systematically review over 700 papers to identify challenges, reveal promising directions, and accelerate development. We propose a taxonomy of UFM architectures based on coupling degree: external service-integrated, modular joint, and end-to-end unified modeling. We analyze encoding/decoding strategies across representations (continuous vs. discrete) and modalities (image, video, audio), discuss the training lifecycle (pre-training, instruction fine-tuning, alignment) with benchmarks for understanding, generation, and mixed tasks, and review applications in robotics, autonomous driving, medicine, and vision, highlighting strengths and weaknesses. Based on our analysis, we summarize trends and discuss defects, such as pure autoregressive/diffusion underperforming hybrid paradigms (which require extra objectives), dual-branch tokenizers introducing redundancy, and gaps in RL reward models, algorithms, and benchmarks. In conclusion, this work aims to guide and inspire further research in building more advanced UFMs. In conclusion, this work aims to be a foundation to inspire further researches in building more unified and capable multimodal AI systems. The project of this paper is available at https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Unified.

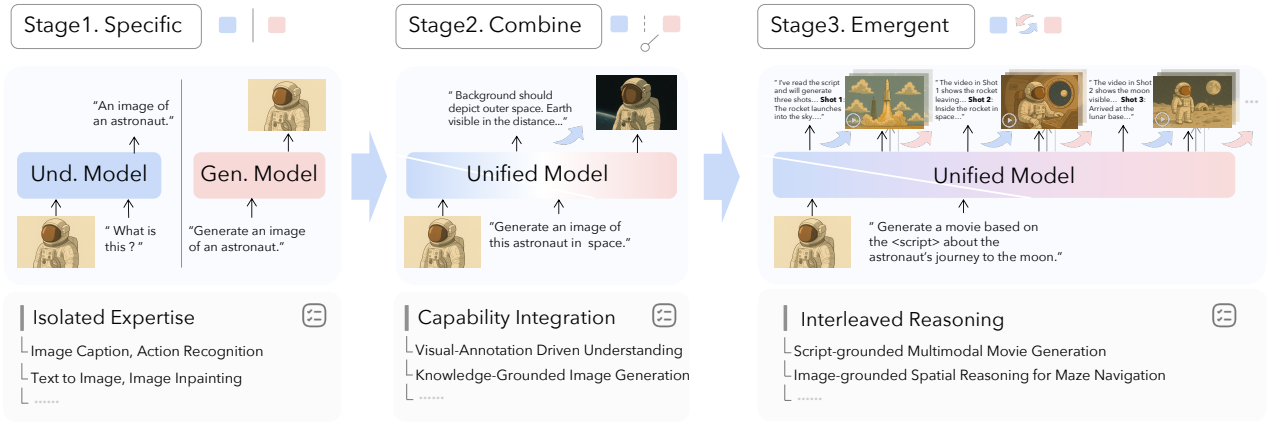

Fig. 1: Evolution of unified multimodal models: From *Specific Stage*, where separate understanding models handle tasks like image captioning and action recognition, and generation models perform text-to-image creation and image inpainting; to *Combine Stage*, enabling Visual-annotation Driven Understanding (e.g., drawing auxiliary lines to better comprehend problems like geometry) and Knowledge-grounded Image Generation (e.g., creating images informed by real-world context); to *Emergent Stage*, as a future vision, facilitating Interleaved Reasoning for complex tasks such as Script-grounded Multimodal Movie Generation (e.g., understanding a complete script and generating its corresponding movie) and Image-grounded Spatial Reasoning for Maze Navigation (e.g., using visuals to reason about paths and environments), which no current work has fully achieved.

Index Terms—Multimodal Large Language Model, Unified Foundation Model, Auto Regressive, Diffusion, Tokenizer

[1] *Nanjing University,* [2] *CASIA,* [3] *PKU,* [4] *NTU*
[*] *Equal Contribution;* [†] *Haochen Tian is the project leader.*
[‡] *Corresponding to : yifanzhang.cs@gmail.com, bradyfu24@gmail.com*

CONTENTS

1 INTRODUCTION

The pursuit of Artificial General Intelligence (AGI) and the development of robust AI systems depend critically on the ability to comprehend and generate multimodal content. For a considerable time, researchers opt to develop generation and understanding independently to focus on breakthroughs within each track. This separation strategy allows for targeted advancements on specific issues, avoids the complexity of joint optimization, and leads to significant progress in low-level tasks (e.g., single-image captioning [1] or unconditional synthesis [2]). However, with breakthrough advancements in understanding and generation, such as the maturation of MLLMs in multimodal perception [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] and the enhancements in high-quality synthesis via diffusion models [13], [14], [15], [16], [17], [18], an increasing number of studies explore the design and training of more challenging and flexible unified models. Recently, breaking down the boundary between understanding and generation to develop a **Unified Foundation Model (UFM)**, capable of seamlessly integrating both, becomes a research hotspot.

The motivation for unifying multimodal understanding and generation is profound, as encapsulated by the famous assertion of physicist Richard Feynman: "What I cannot create, I do not understand." This statement eloquently reveals the inseparable synergy between understanding and creation. We argue that this synergy is not merely a philosophical insight but also provides a core guiding principle for building more powerful artificial intelligence at a technical level.

- **Understanding and Generation as Mutual Reinforcement.** On one hand, deep understanding serves as the foundation for controllable, high-quality generation. For example, when handling a complex instruction like "create a crystal-clear glass chess piece whose surface reflects a burning forest," the model should first accurately comprehend concepts such as "crystal-clear," "glass material," and "reflection," along with the physical and artistic constraints they impose. Without deep semantic understanding, the model, even if it assembles relevant elements, will likely produce a clumsy imitation rather than a faithful representation. On the other hand, generative capability is an indispensable component of complex reasoning, such as drawing auxiliary lines in solving geometry problems or enhancing spatial reasoning and perceptual abilities through "thinking with generated images" [19], [20], [21]. This mutual reinforcement enables unified models to deepen their cognition and reasoning through feedback loops, a powerful capability lacking in models focused solely on understanding or generation.
- **Unified Models for High-Difficulty Real-World Tasks.** Unified models hold irreplaceable value in solving high-difficulty real-world tasks. For instance, generating a short film based on a script requires the model to simultaneously understand narrative structure, visual semantics, and temporal dynamics while iteratively producing coherent scenes. Such high-difficulty tasks inevitably demand seamless collaboration between understanding and generation: the model need deeply comprehend long-context content and generate high-quality multimodal outputs accordingly. Although such tasks are relatively rare in the current research community, we believe that as AI evolves toward more complex real-world applications, increasingly difficult tasks will emerge. UFMs' output forms are more flexible, enabling the execution of more and harder tasks.

Driven by this vision, both academia and industry are actively building UFMs, with a plethora of excellent work emerging. To date, numerous open-source [22], [23], [24], [25], [26] or closed-source [10], [11] UFM efforts exist, and they perform comparably to specialized models in their respective domains (e.g., Qwen 2.5 VL [4], SDXL [27], FLUX.1-dev [14]) for understanding and generation tasks, demonstrating substantial potential. However, open-source models still lag significantly behind closed-source models like GPT-4o [10] and Gemini [11]. The development of the UFM field continues to face many open problems, with no widely recognized technical roadmap achieving the performance levels of closed-source models. For example, modeling approaches remain inconsistent (some works use pure autoregressive frameworks, while others employ autoregressive + diffusion methods); modality tokenizers vary (some rely solely on semantic-level tokenizers like CLIP, while others adopt dual-branch architectures, using pixel-level encoders like VAE for generation and CLIP-like encoders for understanding); and training strategies and data also differ across studies. These issues have yet to reach consensus, limiting the next steps in UFM advancement.

To address this gap, this survey is dedicated to providing a comprehensive and systematic taxonomy and analysis for UFMs. We have collected and organized approximately **700** articles, aiming to establish a detailed classification system for UFMs. As the overview presented in Fig. 2, we hope that by observing and summarizing the evolution of UFMs across dimensions such as modeling approaches, encoding strategies, decoding strategies, training strategies, and evaluation methods, we can address the open questions in UFM model design to the best of our ability, identify current weaknesses, and point out potential development directions. Our main contributions are as follows:

- **A Systematic Taxonomy of Unify Modeling Approaches.** Based on the degree of coupling between the understanding and generation modules, we categorize the architectural paradigms of existing UFMs into three types: *External Service-Integrated Modeling*, *Modular Joint Modeling*, and *End-to-End Unified Modeling*. This provides a clear framework for their design principles.
- **A Panoramic Review of Encoding and Decoding Strategies.** We conduct an exploration of encoding and decoding strategies, which are central to a model's multimodal input and output capabilities. We systematically review mainstream methods based on the form of data representation (continuous vs. discrete) and their application across modalities (image, video, audio).
- **A Holistic Analysis of the Training Lifecycle.** We organize the training strategies and data utilized throughout the entire model lifecycle. Our analysis covers the critical stages of *Pre-training*, *Instruction Fine-tuning*, and *Alignment Fine-tuning*, offering a methodological panorama for constructing and refining UFMs.

- **A Thorough Summary of Benchmarks.** We summarize and organize the existing benchmarks into a clear taxonomy. This covers benchmarks for understanding, generation, and mixed-modality tasks across different modalities, offering a helpful guide for researchers.
- **Applications and Challenges.** We survey the primary application domains where UFMs show great potential, including robotics, auto-driving, world models, medicine, and vision tasks, and summarize the core challenges currently faced.
- **Future Directions.** Looking ahead, promising directions for UFMs include but are not limited to: for modeling strategies, exploring hybrid autoregressive-diffusion modeling and different Mixture of Experts (MoE) structures to better integrate understanding and generation; for tokenizers, scalable unified tokenizers to reduce representation redundancy and enhance generality, and designing efficient video tokenizer with richer semantics for video understanding and generation; for training, designing effective unified reward models, and optimizing understanding and generation capabilities simultaneously in an RL framework remains an open question; finally, benchmarks that directly evaluate mutual promotion between understanding and generation are still lacking.

The remainder of this survey is structured according to the aforementioned contributions. Chapter 3 presents an in-depth discussion of modeling architectures. Chapters 4 and 5 systematically review encoding and decoding strategies, respectively. Chapters 6, 7, and 8 elaborate on methodologies for pre-training, fine-tuning, as well as the data required. Chapter 9 introduces our compiled taxonomy of benchmarks. Chapters 10 and 11 examine applications and future research directions. We summarize some representative open-source methods in Tab. 1, and hope that the survey provides a clear roadmap for newcomers and serves as a valuable reference for researchers in the field.

2 Preliminary

Before the emergence of UFMs, researches on multimodal understanding and multimodal generation progressed independently. These efforts significantly advanced the capabilities of multimodal models in diverse domains, laying the groundwork for the subsequent development of UFMs. In this section, we first provide an overview of multimodal understanding models (Sec. 2.1) and multimodal generation models (Sec. 2.2), establishing the foundational knowledge necessary for understanding UFMs. Subsequently, we formally define UFMs based on their comprehensive task scope (Sec. 2.3), which encompasses both understanding and generation capabilities across multiple modalities.

2.1 Multimodal Understanding Models

Multimodal understanding is foundational in multimodal machine learning, enabling models to perceive and interpret complex visual and textual information. Traditional multimodal understanding tasks include image classification, captioning, visual question answering, etc. Existing approaches are broadly categorized as discriminative or generative. Discriminative models learn decision boundaries to classify multimodal data, while generative models capture joint data distributions, facilitating understanding through generation. We discuss both paradigms below.

2.1.1 Discriminative Models

Given a training set (X, Y), where X denotes the data and Y the labels, discriminative models aim to estimate the conditional probability $p(y|x)$ for label prediction. These models are foundational for tasks such as classification and recognition. Early discriminative models, exemplified by ResNet [28], significantly advanced machine learning's representational capacity. The advent of large-scale pre-trained models further transformed the field. Methods such as MoCo [29] and SimCLR [30] leverage unsupervised pre-training to learn transferable features from unlabeled data, followed by task-specific fine-tuning. More recent models, including CLIP [31], ALIGN [32], SigLip [33], and BiT [34], demonstrate superior generalization and performance across diverse domains. Discriminative models are typically categorized as self-supervised or supervised, depending on the availability of labeled data during training.

Self-supervised Model. Self-supervised learning serves as a foundational paradigm in model training. Its key advantage lies in utilizing supervisory signals derived from the data itself, significantly reducing the need for manually labeled data and enabling large-scale training across various models. Depending on how these signals are generated, self-supervised learning is primarily categorized into contrastive and non-contrastive methods.

Contrastive Learning has gained significant traction in the pre-training of large models, notably with the success of CLIP [31]. It focuses on learning representations by contrasting positive and negative sample pairs, effectively distinguishing between semantically similar and dissimilar instances. The fundamental principle involves maximizing the similarity between positive pairs, which are semantically similar, while minimizing the similarity between negative pairs, which are semantically dissimilar. This is typically achieved using contrastive loss functions. The most widely used is the InfoNCE loss, which is formulated as:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{1}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the embedding vectors of a positive pair of samples, $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ denotes a similarity measure such as cosine similarity, τ is a temperature parameter that adjusts the smoothness of the distribution, and N represents the batch size, including positive and negative samples. By optimizing this loss, contrastive learning captures meaningful patterns, facilitating robust representation learning.

Building on these methods, models such as ALIGN [32], BASIC [35], and Florence [36] have utilized contrastive learning with image-text pairs to enhance their capabilities in image classification and text matching. Additionally, models like OpenCLIP [37] and SigLip [33] are widely employed as image encoders for various tasks. Recently, significant research has focused on extending the contrastive learning paradigm to develop more comprehensive models across various modalities. LanguageBind [38] and ImageBind [39] aim to integrate multiple modalities into a unified

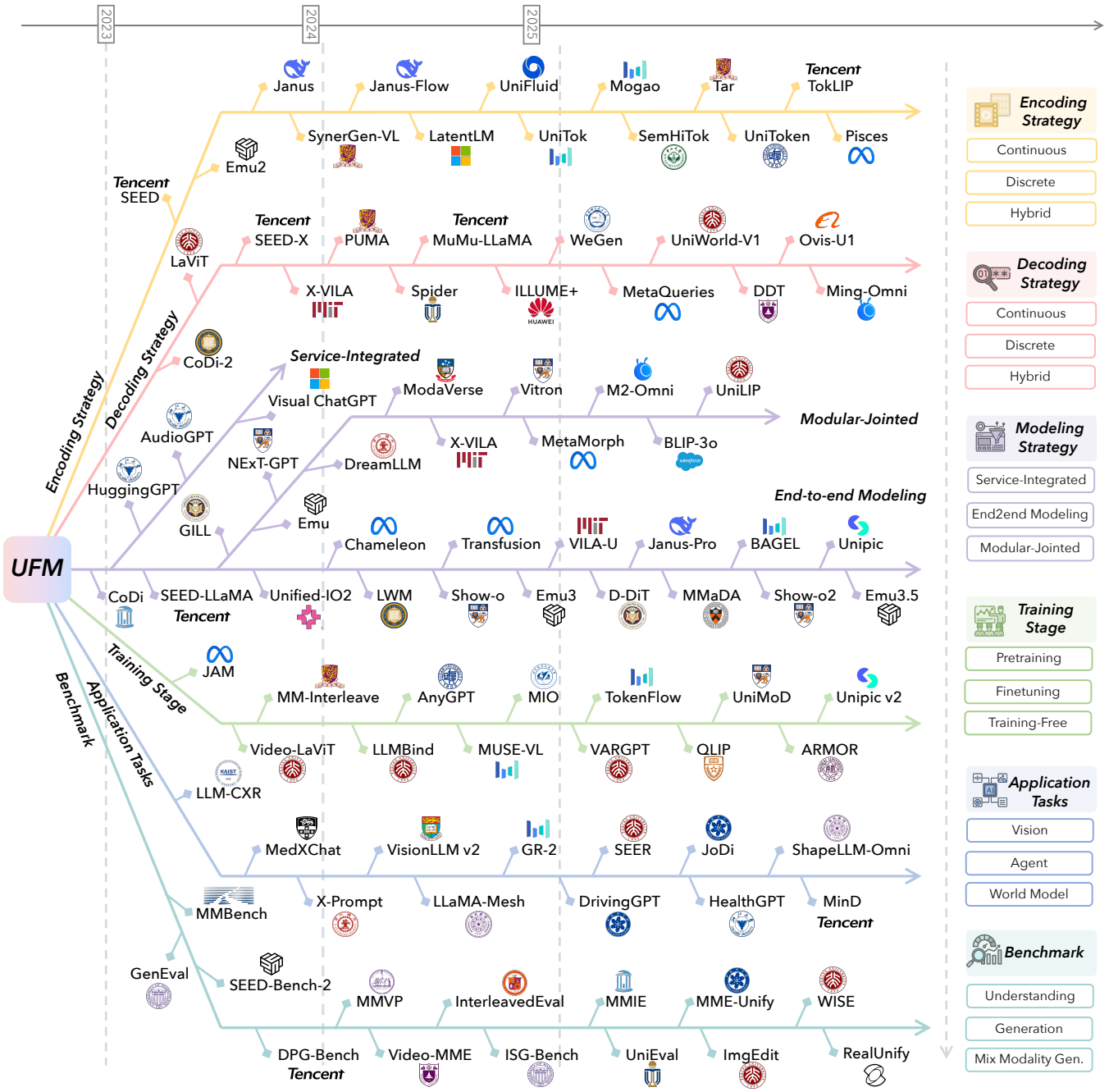

Fig. 2: Overview of the research landscape of UFMs, categorized into six key dimensions: Encoding, Decoding, Modeling, Training, Application, and Benchmark. Representative works are positioned both by category and release year (2023-present), revealing the chronological trajectory of advances across the full technology stack. This organization highlights how prior efforts collectively shape the evolution of the field and indicate emerging trends.

embedding space, where each modality serves as positive samples. This approach facilitates the creation of unified encoders for diverse modalities.

Non-contrastive learning methods differ from contrastive learning primarily in their use of negative samples to construct self-supervised signals. For instance, SimSiam [40] trains models by maximizing the similarity between two augmented views without relying on positive and negative sample pairs. Similarly, DINO [41] involves feeding the same randomly transformed input into distinct student and teacher networks, aiming to align the student network's output with that of the teacher network. This approach exemplifies self-supervised learning without contrastive elements. Additionally, other work [42] also employ non-contrastive self-supervised learning techniques.

Supervised Model. Supervised training is essential in traditional machine learning, where models optimize parameters using labeled data to master specific tasks. Pre-training models on human-labeled datasets can significantly enhance their generalization capabilities. Classic models, such as BiT [34], leverage large-scale classification datasets like JFT [43] and I2E [44] to boost generalization. Despite

challenges in data annotation, supervised models remain well-developed and integral to the training process.

2.1.2 Generative Models

In contrast to discriminative models, generative models aim to learn the joint distribution $P(X, Y)$ rather than directly modeling the conditional probability $p(y|x)$. By capturing the underlying data distribution, generative models enable robust representation learning and facilitate understanding through data synthesis. Foundational approaches in this domain are primarily divided into the masked image modeling (MIM) and autoregressive paradigms.

Masked Image Modeling. The random mask modeling strategy is widely used for pre-training models and is initially demonstrated to be highly effective in the NLP field, as seen in BERT [45] and its subsequent developments. BERT employs random masking for bidirectional language modeling, enabling the model to grasp language meaning and perform comprehension tasks in NLP. Inspired by this approach, BEIT [46] introduced mask prediction training for image encoding to tackle image understanding tasks, while BEVT [47] applied mask modeling to video understanding. To enhance pre-training effectiveness and scalability, MAE [48] is proposed, achieving remarkable results with MAE-style pre-training. By using a high masking ratio and directly predicting pixels, MAE is more scalable and easier to train. VideoMAE [49] further extends this strategy to video understanding. Building on these advancements, SimMIM [50] improves the MIM strategy by replacing the heavy decoder with a single-layer prediction head, enhancing efficiency. Recent research has explored scaling up masked visual representations for large-scale visual pre-training. EVA [51] and EVA02 [52] advance MIM by directly predicting visual features, facilitating large-scale pre-training for vision models. EVA-CLIP [53] integrates this concept to enhance CLIP with extensive pre-training. Collectively, MIM effectively and efficiently builds foundational capabilities for vision models, facilitating downstream applications.

Autoregressive Models. Autoregressive (AR) modeling is a foundational technique extensively employed for data modeling and prediction. In the context of large-scale pre-trained models, AR approaches have become increasingly prominent, particularly following the success of GPT-2 [54]. By sequentially predicting each element conditioned on preceding elements, AR models are well-suited to a wide range of real-world tasks. Recent research has demonstrated the effectiveness of autoregressive models in multimodal understanding. Representative works such as LLaVA [55], LLaVA-1.5 [56], Qwen-VL [57], DeepSeek-VL [58], Qwen2-VL [59], Qwen2-Audio [60], Qwen2.5-Omni [61], and other models [8], [62], [63], [64], [65] have adopted autoregressive frameworks for multimodal understanding. These models leverage AR mechanisms for tasks such as image interpretation, video comprehension, audio analysis, and multimodal input processing. The demonstrated robustness and versatility of autoregressive methods in modeling and interpreting multimodal information have established them as a mainstream paradigm in contemporary model development.

2.2 Multimodal Generation Models

Generative models play a pivotal role in machine learning by enabling the synthesis of diverse and realistic data. Classical approaches, such as energy-based models [66], GANs [67], and autoencoders [68], have demonstrated strong generative capabilities across various domains. Recent advancements in diffusion models [15], flow matching [69], and autoregressive models [54] have further enhanced generative fidelity and versatility, supporting high-quality synthesis across multiple modalities and facilitating unified multimodal integration.

2.2.1 Energy Based models

Energy based models are commonly used methods in statistical physics and have recently received new developments in generative artificial intelligence [70]. These models employ an energy function to represent the probability density of data, utilizing a neural network to minimize the energy associated with the data. The model is defined as follows:

$$p(x) = \frac{e^{-E_\theta(x)}}{Z_\theta}, \tag{2}$$

where E is the energy function, θ is the parameter of the neural network and Z is the normalization constant.

Energy-based models are versatile and widely applied across various tasks, such as synthesizing image [71], and 3D data [72], as well as image recovery [73] and super-resolution [72]. While these models offer stability and flexibility, normalizing the energy function poses significant computational challenges [74].

2.2.2 GAN

Generative Adversarial Networks (GANs) [67] are a successful breakthrough for generative models, opening exciting possibilities for GenAI [75]. It implicitly learns the distribution $q(x)$ of the datasets through the mutual game between the generator module and the discriminator module in the model. The goal of the discriminator is to distinguish between generated data and real data, while the goal of the generator is to generate data that can deceive the discriminator. This process can be expressed as the following two-player minimax game with value function $V(G, D)$.

$$\begin{aligned} \min_G \max_D V(G, D) = E_{x\sim p_x}[logD(x)]+ \\ E_{z\sim p_z}[log(1 - D(G(z)))], \end{aligned} \tag{3}$$

where G and D represent the generator and discriminator, respectively, and p_x and p_z represent the input data distribution and noise distribution, respectively.

Initially, GAN is introduced to generate new images from the MNIST dataset. Subsequently, GAN-based models such as SRGAN [76] and SR [77] are developed to enhance image resolution. Other studies [78], [79], [80], [81] have integrated GANs for tasks like image-to-image translation and style transfer.

To enhance the performance and stability of GANs, researchers have proposed numerous strategies [82], [83], [84], such as refining loss functions [85], [86], [87] and modifying network architectures [88], [89], [90]. Despite these improvements, GANs are often plagued by training instability [67], [91], as the adversarial optimization between the generator

and discriminator can lead to convergence difficulties [74]. As a result, alternative generative models, including diffusion models [15] and flow matching [69], have emerged to address these limitations and further advance the field.

2.2.3 AE & VAE

Autoencoders [68] are typically a type of algorithm proposed by Rumelhart [92] for reconstructing a set of input observations. Autoencoder mainly consists of three parts: an encoder, a latent feature representation, and a decoder [68]. The encoder and decoder are generally composed of neural networks, denoted as E and D. Therefore, for an input x, the output of the model can be represented as $\hat{x} = D(E(x))$. The latent features are represented by the intermediate variable $h = E(x)$. So the training objective for the autoencoder model can be expressed as below:

$$\arg\min_{D,E} < \delta(x, D(E(x))) >, \tag{4}$$

where δ represents the distance between x *and* $\hat{x}$, $< . >$ indicate the average value of the observed variable.

Autoencoder-based reconstruction models can effectively capture multi-modal information. However, their use in generation tasks was initially limited because they do not inherently produce new data [75]. The introduction of VAEs [93] addressed this limitation by replacing fixed latent variables with numerical distributions, allowing latent variables to be sampled for data generation. In training, VAEs incorporate both classical reconstruction loss and KL divergence to ensure the latent variables maintain a normal distribution. This approach effectively models multi-modal information, and many subsequent methods have utilized VAE modules to interpret input data.

Building on VAEs, VQVAE [94] introduces a codebook mechanism that maps encoding vectors into a discrete latent space, enabling more structured and efficient representations for generative tasks. By leveraging an embedding space and employing PixelCNN [95] as the generative prior, VQVAE facilitates high-quality data synthesis. Subsequently, VQGAN [96] extends this framework by replacing PixelCNN with a transformer-based autoregressive decoder and incorporating adversarial loss through PatchGAN [78] discriminators, thereby achieving superior content generation fidelity. These AE-based methods, particularly those utilizing vector quantization and adversarial training, are not only effective for a wide range of generation tasks but also serve as foundational components in modern generative models [97], [98] for both encoding and decoding, as exemplified by VQVAE and VQGAN. However, in the recent work, RAE [18] has attracted a lot of attention, showing a trend of replacing VAE, but still using the encoder-decoder paradigm also proves the reliability of this paradigm.

2.2.4 Diffusion Models

Diffusion models achieve impressive generation capabilities through a forward-reverse strategy. The forward process incrementally adds Gaussian noise to the initial data x_0 over timesteps T, resulting in x_T. Subsequently, the reverse process is trained to learn how to denoise, reconstructing data samples from random Gaussian noise. In the forward process, given the data distribution $x_0 \sim q(x_0)$, the data x_t at each step t is represented as follows:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{5}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{6}$$

where β_t denotes the variance schedule of the noise to control the noise we add to the original data. In the denoising process, the model takes a sample from a gaussian noise as input and learns to reconstruct the original data. The denoise process is based on the distribution

$$p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I}), \tag{7}$$

as the forward process will give by

$$q(x_T) = \mathcal{N}(x_T; 0, \mathbf{I}). \tag{8}$$

Then for each denoising step, a learnable transition kernel of x_t to x_{t-1} is introduced:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{9}$$

where θ denotes the parameters of the network and the mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$ are parameterized by the network. Following this, the network takes a noise sampled from $p(x_T)$ and iteratively performing the denoise steps from t to $t-1$ until $t = 1$. The widely-used training objective of diffusion models can be represented as below:

$$E_{t,x_0,\epsilon}\left[\lambda(t)\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \tag{10}$$

where $\lambda(t)$ is a positive weighting function, x_t is computed from x_0 and ϵ as described before, and ϵ_θ is a deep neural network with parameter θ that predicts the noise vector ϵ given x_t and t.

When diffusion models are first introduced, they commonly apply the diffusion process in pixel space, leading to works like GLIDE [99] and Imagen [100]. While this approach is straightforward, it suffers from high training costs and inference latency, prompting the development of latent diffusion models. Latent diffusion models perform the diffusion process in latent space using a pretrained VAE, reducing training costs while maintaining high image generation quality. This approach has become mainstream, inspiring works such as SD-1.5 [15] and SD-XL [27]. Advanced research explores methods to further accelerate the diffusion process, such as Consistency Models [101], which achieve high-quality image generation within a single step.

In addition to training methodologies, the architectural design of diffusion models plays a pivotal role in their performance. Traditionally, U-Net [102] architectures have been employed to encode and reconstruct fine-grained image details. However, the emergence of transformer-based architectures has introduced new possibilities. Notably, DiT [16] replaces the U-Net backbone with a transformer-based design, representing images as sequences of tokens and incorporating auxiliary inputs such as timesteps and control signals. These tokens are processed through multiple layers of transformer blocks, leveraging attention mechanisms to preserve image fidelity and effectively integrate conditioning information. This transformer-based approach not only enhances the scalability of diffusion models but also facilitates

the training of larger and more expressive generative models. Recent advancements, including SD-3.0 [103], SiT [104], and Sora [105], have adopted and extended this architectural paradigm, underscoring its significance in advancing the capabilities of diffusion-based generative models.

2.2.5 Normalizing Flow

The normalizing flow model follows the principle of modeling data distributions in generative models by learning sample distributions to generate probabilities. Its most notable feature is its reversibility, constructing bijective transformations [74], [106]. In normalizing flow, the model is considered as a transformation f, mapping the input x to the output $o = f(x)$, while transforming the original distribution p_x to the target distribution p_o. The probability density of x can be expressed as follows:

$$p_x(x,\theta) = p_o(f^{-1}(x)) \left| \det \frac{\delta f^{-1}(x)}{\delta x} \right|, \tag{11}$$

which is derived from the rule of change of variables [107]. Based on the predicted probability density, the model can be trained directly by minimizing the negative log-likelihood:

$$L = -\log p_x(x,\theta). \tag{12}$$

There are also related approaches [106] that formulate training objectives by introducing additional constraints, such as the Kullback–Leibler (KL) divergence, which serve not only as a means of regularization but also as a way to guide the flow toward more stable densities and improve generalization across different data distributions.

Normalizing flow is used primarily for image generation, with models such as NICE [108], Real NVP [109], and Glow [110]. It also finds applications in text modeling [111] and audio modeling [112]. However, normalizing flow is not a popular method previously, possibly due to the constraints imposed by the need for reversibility and easily computable Jacobian determinants, which limit the types of transformations that can be employed. Recent advances have sparked renewed interest in normalizing flows. Flow matching [69] simplifies the training objectives of continuous normalizing flows, significantly reducing training challenges while maintaining impressive performance, potentially paving the way for more powerful models.

2.2.6 Autoregressive models

Autoregressive models have increasingly become the dominant framework for artificial intelligence, particularly in multi-modal understanding [113]. Their importance in multi-modal generation is also growing. They generate subsequent elements based on preceding ones, as shown below:

$$p(x) = \prod_{i=1}^{N} p(x_i | x_1, x_2, ..., x_{i-1}; \theta), \tag{13}$$

where $p(x_i|x_1, x_2, ..., x_{i-1}; \theta)$ refers to the conditional probability of predicting the current element x_i based on the previous elements, and θ is the model parameter.

Then, the training objective is to minimize the likelihood loss of the sum of the negative logarithms of the predicted probabilities of all elements, which can be shown as below:

$$L(\theta) = -\sum_{i=1}^{N} \log p(x_i | x_1, x_2, ..., x_{i-1}; \theta). \tag{14}$$

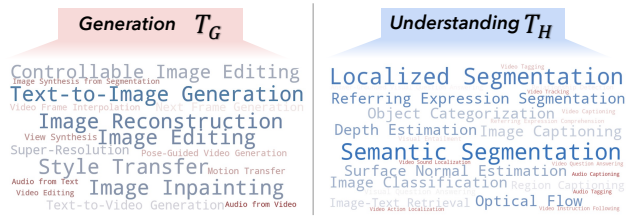

Fig. 3: An illustration of the unified task set. The unified task set is defined as a collection of tasks that includes both understanding tasks (T_U) and generation tasks (T_G).

Recently, numerous studies have focused on autoregressive models for multi-modal generation. PixelRNN [114] and PixelCNN [95] are classic models that generate images pixel by pixel, constructing complete images from known pixels and offering innovative approaches to image generation tasks. Autoregressive models have now become a mainstream architecture in generative artificial intelligence. Recent research [115], [116], [117] continues to build on these architectures, and their use in multi-modal models is increasingly prevalent.

The primary advantage of autoregressive models is their alignment with real-world generation processes, making them suitable for producing long sequences with enhanced performance. However, their dependence on preceding elements limits parallelism, leading to reduced output efficiency and potential error accumulation [118].

2.3 Unified Task Formulation

Although unified foundation models have achieved remarkable progress, the precise definition of a "unified foundation model" remains ambiguous. To provide a rigorous research framework, this paper first formalizes the associated tasks.

Given an arbitrary input data x, its corresponding task type can be denoted as:

$$t_x := \text{the task type corresponding to data } x. \tag{15}$$

The unified foundation models discussed in this paper are those capable of handling both multimodal understanding and generation tasks. To this end, we first define the sets of understanding and generation tasks:

$$\begin{cases} T_U := \{t \mid t \text{ is an understanding task}\} \\ T_G := \{t \mid t \text{ is a generative task}\} \end{cases}. \tag{16}$$

Here, T_U and T_G represent the sets of all generative and understanding tasks, respectively. Examples of these tasks are listed in the Appendix. It is worth noting that both T_U and T_G are open sets; the tasks listed in this paper constitute only a currently known subset, which can be further expanded by the research community in the future.

Building upon these definitions, we assert that the task set $UniSet$ addressed by a unified foundation model should encompass both understanding and generation tasks. Consequently, the collection of all admissible $UniSet$s, denoted as $PowerUniSet$, can be formally expressed as follows:

$$PowerUniSet = 2^{T_U \cup T_G} - 2^{T_U} - 2^{T_G}, \tag{17}$$

where $PowerUniSet$ denotes the set of all non-empty subsets of $T_G \cup T_U$ that contain at least one generative task and one understanding task as in Fig. 3. This guarantees that a unified foundation model need to be capable of handling both types of tasks, rather than specializing in only one.

Accordingly, we formally define a unified foundation model (UFM) as follows:

Definition. A unified foundation model (UFM) is characterized by the following properties:

- There exists a task set $I \in PowerUniSet$.
- For any task $t \in I$, the model can process the corresponding input data and produce appropriate outputs.
- The choice of I determines the extent of the model's unified capabilities. A larger I indicates broader task coverage and stronger model capability.
- In this paper, we can refer to the unified foundation model as the unified model or the UFM.

It is important to note that this paper primarily concentrates on UFMs incorporating vision-related modalities. Models that exclusively possess audio generation capabilities [119], [120], [121] fall beyond the scope of this investigation and are not extensively discussed herein. For a comprehensive analysis of such models, readers are referred to the relevant sections of the omni-modal model survey [122].

Furthermore, due to the scale of models and the complexity of data, existing models commonly adopt phased training strategies, such as the "pre-training and fine-tuning" paradigm. However, in the domain of large multimodal models, the boundary between pre-training and fine-tuning has become blurred, hindering fair comparative studies. Proceeding from the objective of unified models, we define the process of a model learning to handle both understanding and generation tasks simultaneously as **Unified Pre-training (UP)**. Any subsequent training to enhance performance on specific tasks is then defined as fine-tuning.

Formally, a model M initialized from a LLM inherently possesses only multimodal understanding capabilities. In this scenario, the model is restricted to processing inputs associated with understanding tasks, and its outputs are exclusively determined by the nature of these tasks:

$$output = M(x), \quad x \in \{x \mid t_x \in T_U\}. \tag{18}$$

To transform it into a unified model, it should undergo unified pre-training to acquire the ability to handle generative tasks. The pre-trained model $M_{pretrained}$ should satisfy:

$$output = M_{pretrained}(x), \tag{19}$$

where $\forall x$ s.t. $t_x \in I$, for some $I \in PowerUniSet$.

In this manner, we formally define the unified pre-training process from the perspective of task coverage. Fine-tuning, then, builds upon this foundation, utilizing specific data (such as instruction-following data) to further optimize model parameters for enhanced performance.

3 Modeling

Due to the distinct representational forms, generation mechanisms, and task characteristics across modalities, designing effective modeling approaches represents a fundamental challenge in developing UFMs. Unlike conventional multimodal understanding models or generation models, UFMs should jointly optimize understanding and generation objectives, which frequently results in conflicting requirements, such as learning high-level semantic representations while preserving low-level textural details. Moreover, UFMs often necessitate the integration of disparate modeling paradigms (e.g., combining autoregressive and diffusion-based frameworks), introducing substantial complexity in training and inference. The modeling strategy thus critically determines the balance between theoretical rigor and practical implementation, fundamentally shaping the trajectory of UFM development in both research and deployment contexts.

Based on the coupling mechanisms of different modeling approaches, we classify current UFMs into three categories: **External Expert Integration Modeling** (Sec. 3.1), **Modular Joint Modeling** (Sec. 3.2), and **End-to-End Unified Modeling** (Sec. 3.3). This classification is made according to the degree of coupling, the dependence on external generation modules, and the uniformity of the generation process. It serves to analyze the differences and implications of various modeling strategies with respect to system architecture, capability integration, inference efficiency, and other critical aspects, such as scalability and resource demands.

3.1 External Expert Integration Modeling

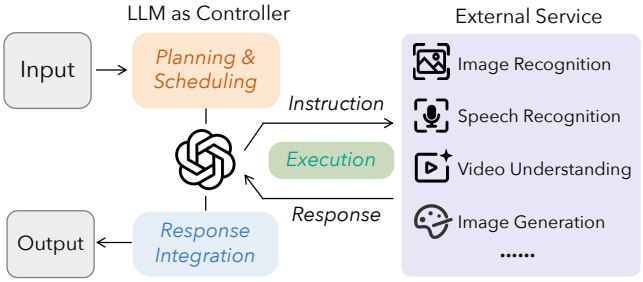

Fig. 4: External Expert Integration Modeling. The overall process consists of three steps: (1) task planning and scheduling, (2) task execution, and (3) response post-processing and integration.

(1) Definition

Given the strong capabilities of LLMs in understanding, planning, and instruction following across natural language processing tasks, some studies [125], [126], [120], [183] adopt LLMs as the central decision-making module. These models interact with external expert models or services through natural language interfaces to achieve multimodal understanding and generation. We refer to this modeling paradigm as "External Expert Integration Modeling".

As shown in Fig. 4, the core idea of this modeling strategy is to leverage the LLM's strong capabilities in contextual understanding, task planning, and unified language-based interfacing, placing it at the center of the system as an orchestrator or controller. Under this approach, the LLM does not directly perform tasks such as image recognition, speech processing, or image generation. Instead, upon receiving user instructions that may involve multimodal information such as speech, images, or videos, it analyzes the user's intent and task objectives, autonomously determines the required processing steps and toolchain combinations, and

TABLE 1: Summary of open-source methods. Modality represents the output modality. Paradigm includes autoregressive (AR), diffusion (Diff), and external expert integration (INTEG), with † indicating external generation modules opposed to end-to-end modeling. Training encompasses unified pre-training (UP), supervised fine-tuning (ST), alignment fine-tuning (AT) and training free (-). GitHub-Star counts are as of September 2025. Some high-impact open-source methods are presented in the main table below, with comprehensive compilation provided in the appendix.

#	Model	Venue	Date	Backbone	Modality	Paradigm	Training	GitHub-Star
1	OFA [123]	ICML	2022.02	S2S Transformer	image	AR	UP+ST	2.5k
2	UNIFIED-IO [97]	ICLR	2022.06	S2S Transformer	image	AR	UP+ST	228
3	Versatile Diffusion [124]	ICCV	2022.11	UNet	image	Diff	UP+ST	1.3k
4	Visual ChatGPT [125]	arXiv	2023.03	Text-davinci-003	image	INTEG†	-	34.4k
5	HuggingGPT [126]	NeurIPS	2023.03	GPT-3.5	image,audio,video	INTEG†	-	24.3k
6	UniDiffuser [127]	ICML	2023.03	GPT-2	image	Diff	UP+ST	1.4k
7	CoDi [128]	NeurIPS	2023.05	LDM	image,audio	Diff	UP+ST	1.7k
8	Emu [129]	ICLR	2023.07	LLaMA	image	AR†	UP+ST	1.7k
9	SEED [130]	arXiv	2023.07	OPT	image	AR†	UP+ST	621
10	NExT-GPT [131]	ICML	2023.09	Vicuna	image,audio,video	AR†	UP+ST	3.6k
11	DREAMLLM [132]	ICLR	2023.09	Vicuna	image	AR†	UP+ST	455
12	LaVIT [133]	ICLR	2023.09	LLaMA	image	AR†	UP	590
13	Minigpt-5 [134]	arXiv	2023.10	Vicuna	image	AR†	UP+ST	861
14	SEED-LLaMA [135]	ICLR	2023.10	Vicuna	image	AR†	UP+ST	621
15	CoDi-2 [136]	CVPR	2023.11	LLaMa-2	image,audio,video	AR†	UP+ST	1.7k
16	Emu2 [137]	CVPR	2023.12	LLaMA	image,video	AR†	UP+ST	1.7k
17	Unified-IO 2 [138]	CVPR	2023.12	S2S Transformer	image,audio	AR	UP+ST	622
18	MM-Interleaved [139]	arXiv	2024.01	Vicuna	image	AR†	UP+ST	237
19	AnyGPT [140]	ACL	2024.02	LLaMA-2	image,audio,video	AR†	UP+ST	858
20	Video-LaVIT [141]	ICML	2024.02	LLaVA-1.5	video	AR†	UP+ST	590
21	LWM [142]	ICLR	2024.02	LLaMa-2	image	AR	UP+ST	7.3k
22	Mini-Gemini [143]	arXiv	2024.03	Nous-Hermes-2-Yi	image	AR†	UP+ST	3.3k
23	Seed-X [144]	arXiv	2024.04	LLaMa-2	image	AR†	UP+ST	533
24	Chameleon [145]	arXiv	2024.05	LLaMa-2	image	AR	UP+ST	2k
25	TextHarmony [146]	NeurIPS	2024.07	MM-Interleaved	image	AR†	UP+ST	130
26	ANOLE [147]	arXiv	2024.07	Chameleon	image	AR	ST	796
27	Lumina-mGPT [148]	arXiv	2024.08	Chameleon	image	AR	UP+ST	622
28	Show-o [116]	ICLR	2024.09	Phi-1.5	image	AR+Diff	UP+ST	1.7k
29	VILA-U [149]	arXiv	2024.09	LLaMA-2	image,video	AR	UP	382
30	Emu3 [25]	arXiv	2024.09	LLaMa-2	image,audio	AR	UP+ST+AT	2.2k
31	Janus [150]	CVPR	2024.10	DeepSeek-LLM	image	AR	UP+ST	17.5k
32	PUMA [151]	arXiv	2024.10	LLaMA-3	image	AR	UP+ST	130
33	JanusFlow [152]	CVPR	2024.11	DeepSeek-LLM	image	AR+Flow	UP+ST	17.5k
34	MoT [153]	ICLR	2024.11	Chameleon	image,audio	AR	UP+ST	93
35	TokenFlow [154]	CVPR	2024.12	Qwen-2.5	image	AR	UP+ST	373
36	MetaMorph [155]	arXiv	2024.12	LLaMA-3.1	image,video	AR†	UP+ST	208
37	Liquid [156]	arXiv	2024.12	LLaMa-2	image	AR	UP+ST	612
38	MuMu-LLaMA [157]	arXiv	2024.12	LLaMA	video	AR†	UP+ST	500
39	Janus-Pro [26]	arXiv	2025.01	DeepSeek-LLM	image	AR	UP+ST	17.5k
40	VARGPT [158]	arXiv	2025.01	Vicuna-1.5	image	AR	UP+ST	347
41	UniTok [159]	arXiv	2025.02	LLaMa-2	image	AR	UP+ST	392
42	QLIP [160]	arXiv	2025.02	Vicuna-1.5	image	AR	UP+ST	83
43	Harmon [161]	arXiv	2025.03	Qwen-2.5	image	AR	UP+ST	156
44	UniDisc [162]	arXiv	2025.03	S2S Transformer	image	Diff	UP	119
45	OmniMamba [163]	arXiv	2025.03	Mamba-2	image	AR	UP+ST	137
46	ILLUME+ [164]	arXiv	2025.04	Qwen-2.5	image	AR†	UP+ST	118
47	UniToken [165]	CVPR	2025.04	Chameleon	image	AR	UP+ST	88
48	MetaQueries [166]	arXiv	2025.04	Qwen-2.5-VL	image	AR†	UP+ST	222
49	VARGPT-v1.1 [167]	arXiv	2025.04	Qwen-2	image	AR	UP+ST+AT	262
50	Nexus-Gen [168]	arXiv	2025.04	Qwen-2.5-VL-Instruct	image	AR†	UP+ST	266
51	BAGEL [22]	arXiv	2025.05	Qwen-2.5	image,video	AR+Flow	UP+ST	4.9k
52	BLIP3-o [23]	arXiv	2025.05	Qwen-2.5-VL-Instruct	image	AR†	UP+ST	1.4k
53	Ming-Lite-Uni [169]	arXiv	2025.05	LLaMA-3	image	AR†	UP	446
54	TokLIP [170]	arXiv	2025.05	Qwen-2.5	image	AR	UP+ST	186
55	UniCTokens [171]	arXiv	2025.05	Show-o	image	AR+Diff	UP+ST	121
56	MMaDA [172]	arXiv	2025.05	LLaDA-Instruct	image	Diff	UP+ST+AT	1.3k
57	OpenUni [173]	arXiv	2025.05	InternVL-3	image	AR†	UP+ST	149
58	Ming-Omni [174]	arXiv	2025.06	Qwen-2.5-VL	image,audio	AR†	UP+ST	446
59	Show-o2 [175]	arXiv	2025.06	Qwen-2.5	image,video	AR	UP+ST	1.7k
60	OmniGen2 [176]	arXiv	2025.06	Qwen-2.5-VL	image	AR†	UP+ST	3.8k
61	Ovis-U1 [177]	arXiv	2025.06	Qwen-3	image	AR†	UP+ST	404
62	Janus-4o [178]	arXiv	2025.06	Janus-Pro	image	AR	ST	259
63	X-Omni [179]	arXiv	2025.07	Qwen-2.5	image	AR†	UP+ST+AT	373
64	Omni-Video [180]	arXiv	2025.07	VILA	image,video	AR†	UP+ST	44
65	UniPic [181]	arXiv	2025.08	Qwen-2.5	image	AR	UP+ST+AT	785
66	UniPic-2.0 [182]	arXiv	2025.09	Qwen-2.5-VL	image	AR†	UP+ST+AT	785

constructs formatted prompts in natural language to invoke external expert models for tasks such as image recognition, speech recognition, image generation, and speech synthesis. The LLM then gathers and integrates the results from these expert models and generates the final output.

The design and implementation of External Expert Integration Modeling typically involve three key components. First, task planning and orchestration, where the LLM interprets user intent, decomposes the task into subtasks, and generates structured control instructions to invoke external modules. This process, often driven by prompt-based natural language guidance, requires: (1) task decomposition, which involves breaking down complex inputs into manageable subtasks while preserving the correct input-output dependencies and execution order; (2) selection of appropriate external expert models (e.g., speech recognition, image generation) and modality planning to ensure coherent data flow; and (3) generation of structured prompts specifying task types and input parameters. Second is the execution of external expert models, wherein the LLM uses the generated prompts to call the specific external expert model. These expert models should adhere to a unified interface specification to ensure reliable interaction with the LLM. Third, post-processing and integration of results, which involves not merely aggregating outputs but aligning them with the user's intent and task objectives. In cases where the required functionality exceeds current external expert model's capabilities, the model should be able to provide appropriate fallback responses or guidance.

(2) Relevant Works

In practical implementations, this modeling approach demonstrates notable flexibility and scalability. For instance, Visual ChatGPT [125] employs a Prompt Manager to handle the input-output interfaces of multiple visual foundation models, enabling ChatGPT to perform visual tasks such as visual question answering and image generation. HuggingGPT [126] builds on this concept by developing a comprehensive expert model orchestration system, allowing the LLM to dynamically integrate models from the HuggingFace community to solve complex tasks, with dedicated optimizations for task scheduling. AudioGPT [120] extends the service invocation mechanism to the audio domain, supporting speech recognition, audio editing, and synthesis. In contrast to the previous training-free methods, SwitchGPT [183] enhances the LLM's ability to manage modality transformation tasks through lightweight instruction tuning for modality alignment, without requiring extensive retraining, enabling more flexible use of external expert models and further raising the capability ceiling.

(3) Pros and Cons

External Expert Integration Modeling offers notable advantages in its simplicity and low resource requirements, enabling multimodal understanding and generation without the need for extensive training or large-scale datasets. Moreover, its modular design decouples the LLM from various expert models, allowing components to interact through standardized interfaces. This facilitates flexible replacement and extension of external expert modules, making it well-suited for diverse tasks and application scenarios.

This modeling approach also presents obvious limitations. First, as task execution heavily relies on external expert models, effectively utilizing and optimizing the performance of external expert models requires additional development and debugging efforts to ensure seamless integration and minimize potential bottlenecks. Second, the potential requirement for multiple external calls during task execution can negatively impact the system's overall efficiency. This may also increase the risk of cumulative errors and information loss. Third, when the system involves multiple external models or services, there may be security risks, such as information leakage, especially when invoking remote services that cannot be deployed locally.

Overall, External Expert Integration Modeling serves as an early-stage solution for rapidly enabling multimodal understanding and generation, making it more suitable for prototyping and engineering-oriented implementations.

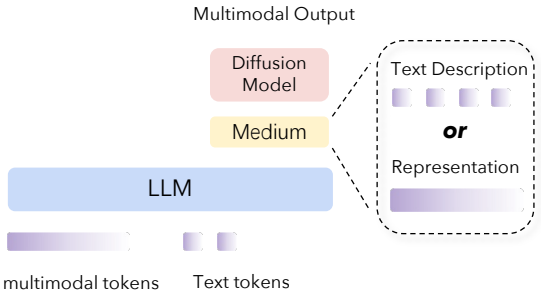

Fig. 5: Modular Joint Modeling. The generation of multimodal content typically requires invoking external generative models (e.g., diffusion models).

3.2 Modular Joint Modeling

Modular Joint Modeling is a modeling paradigm that seeks a flexible connection between understanding and generation capabilities. At its core, this approach adopts a modular architecture within multimodal systems by integrating independent generation modules alongside an LLM backbone, thereby enabling the flexible composition and execution of multimodal tasks. Specifically, the LLM serves as the primary processor responsible for input understanding and contextual modeling. When tasked with generating non-text modalities such as images or video, it produces either descriptive outputs or intermediate representations that guide external generation modules to produce the final target modality content. These external modules are typically well-trained, modality-specific generators such as diffusion models, which possess strong generative capabilities and can significantly enhance the quality of multimodal generation.

Fig. 5 illustrates the overall architecture of Modular Joint Modeling. To provide a more detailed characterization of the internal mechanisms underlying this modeling approach, it can be further divided into two categories based on the generation conditions. The first category is **Prompt-Mediated Modeling** (Sec. 3.2.1), which generates natural language instructions to guide external models in producing content such as images or videos. The second category is **Representation-Mediated Modeling** (Sec. 3.2.2), which conditions external generation modules on intermediate representations. Although these two approaches

differ in representational form and interface design, both adhere to a modular architecture centered on the principle of "understanding-mediation-generation".

3.2.1 Prompt-Mediated

(1) Definition

Prompt-Mediated Modeling establishes a streamlined and extensible pathway by leveraging natural language prompts to bridge the understanding and generation subsystems. The core idea is that, when the model is required to generate non-textual modalities such as images or videos during task execution, it generates a descriptive text prompt, which then drives an external multimodal generation model (e.g., a text-to-image or text-to-video generator) to produce the final output. This approach effectively decouples language understanding from modality-specific generation, enabling more flexible integration of various pretrained generators and significantly reducing both training costs and system complexity. Prompt-Mediated Modeling inherits the paradigm of using natural language as a unified interaction interface and shares conceptual similarities with models that operate through external expert models invocation.

At the implementation level, Prompt-Mediated Modeling typically employs lightweight alignment strategies that combine existing LLMs and generative models. A common approach involves instruction tuning, which teaches the language model to generate effective prompts in response to input tasks, thereby facilitating the connection between understanding and generation. Since it avoids explicit alignment between the feature spaces of the language and generation models, this method not only reduces computational overhead during training but also improves data efficiency.

(2) Relevant Works

In recent years, a variety of representative studies have demonstrated the effectiveness and practicality of Prompt-Mediated Modeling for UFMs, while also showcasing the diversity of implementation paradigms.

Divter [184] is among the earliest attempts to integrate dialogue systems with image generation models. It requires the model to generate not only textual responses but also images conditioned on the dialogue context. Divter's key idea is to decouple text and image generation by leveraging a pretrained language model for textual responses and a separate image generation model for text-to-image translation, with joint training achieved using a limited amount of data. This decoupled design has inspired many subsequent works. Based on this framework, TIGER [185] introduces a response modal predictor that determines whether the next response should be textual or visual, based on contextual cues. This approach enhances coherence in Divter and improves the overall user experience.

Building further, EasyGen [186] connects BiDiffuser and an LLM to enable bidirectional generation between text and images. Unlike Divter [184], which does not support visual understanding, EasyGen can also interpret and encode visual inputs, significantly enhancing its applicability. To further improve handling of high-resolution visual inputs, Mini-Gemini [143] adopts a dual-vision encoder architecture, where a low-resolution encoder supplies visual queries, and a high-resolution encoder provides keys and values, coordinated through attention mechanisms. To manage computational cost, a token mining strategy is employed to prevent excessive visual token expansion while preserving high-fidelity perception.

Modality Integration. Due to the flexibility of Prompt-Mediated Modeling, the framework can be readily integrated with models such as text-to-video and text-to-audio generators, thereby enabling seamless extension of modality coverage within this modeling paradigm.

In the video domain, GPT4Video [187] proposes a parameter-efficient framework that integrates video understanding and generation without modifying model weights. It incorporates a safety detection mechanism through fine-tuning using pure text, effectively mitigating harmful content generation risks and enhancing system security and compliance. In addition to video, ModaVerse [188] further extends its modality coverage to include audio. Inspired by HuggingGPT [126], ModaVerse proposes an "adaptor + agent" architecture, where a lightweight input-side adaptor handles multimodal understanding, and the output-side LLM-as-agent delivers natural language prompts to invoke external generators. Additionally, ModaVerse introduces an I/O alignment strategy along with a dedicated instruction dataset, substantially improving control precision and flexibility over external modalities.

Hybrid Interface. It is worth noting that some works have begun to explore modeling approaches that go beyond a single prompt form, aiming to combine natural language prompts with intermediate representations for more flexible and accurate generation control.

For instance, LLMBind [189], which adopts a Mixture-of-Experts (MoE) architecture, focuses on enhancing the model's ability to handle multiple tasks. By introducing task-specific tokens, LLMBind learns to address different tasks for various modalities and invokes the corresponding expert modules to complete them. This approach improves both the updateability and expandability of the model. In addition to generating `Task-Prompt Tokens` in natural language form, LLMBind also generates `Semantic-Embedding Tokens` as intermediate representations. These tokens contain rich semantic information, which helps the model perform tasks such as image segmentation and detection better. Further extending this concept, Spider [190] expands the UFM's generative capabilities from "Text + X" (such as Text + Image or Audio or Video) to "Text + Xs" (i.e., Text + Image and Audio and Video), enabling a more comprehensive multimodal generative interaction that significantly enhances the user experience. However, the generation of multiple modalities introduces additional complexity. To address this issue, Spider introduces a "Decoders-Controller" that effectively manages multiple decoders. By combining the text prompt with auxiliary modality prompts, it extracts rich modal information and decodes it. VITRON [191] adopts a similar design, using natural language instructions to precisely identify the target generation module and combining it with continuous visual embeddings. This enhances the richness and accuracy of modality feature transmission. In more complex multimodal generation scenarios, M2-Omni [192] further proposes a multi-stage training strategy that progressively aligns text, image, and audio content. Notably,

M2-Omni adopts different strategies for modality-specific generation: for images, it leverages the generated natural language description to invoke Stable Diffusion [15]; for audio, M2-Omni directly inputs discrete audio tokens into a dedicated generation module, CosyVoice [193], which is specifically designed for high-quality speech synthesis.

These works not only expand the expressive boundaries of Prompt-Mediated Modeling but also provide viable pathways toward achieving more fine-grained and consistent multimodal understanding and generation.

(3) Pros and Cons

The prominent advantage of Prompt-Mediated Modeling lies in its architectural simplicity and high scalability. By using natural language as a bridge, this approach avoids the cumbersome intermediate feature alignment mechanisms, maintaining a loosely coupled relationship between the language model and the generation module. This facilitates the integration of higher-performance generative models while keeping maintenance and upgrade costs low. Additionally, since the intermediate outputs exist in textual form, this approach offers excellent human interpretability and interactivity. Users can manipulate the prompts to control the generation results, thereby enhancing model controllability. This modeling approach also naturally adapts to optimization strategies based on human preference alignment, making it easier to introduce safety control mechanisms.

Although Prompt-Mediated Modeling demonstrates significant flexibility and practicality in practice, its generation quality is largely constrained by the expressiveness of natural language prompts. Due to the abstract and ambiguous nature of language, prompts often struggle to fully convey fine-grained semantic information, such as image structure or action sequences. This may result in missing information or unstable performance when external generative modules handle complex tasks. Moreover, for scenarios that require precise control over generation details, such as high-precision medical image analysis or fine-grained control of character behavior, prompt-mediated approaches may fail to meet the high demands for expression accuracy, generation consistency, and contextual relevance.

Therefore, while this method offers clear advantages in terms of generalizability and rapid integration, it still faces challenges in high-precision, multidimensional controllable generation scenarios, which may need to be supplemented and enhanced by stronger modality alignment mechanisms.

3.2.2 Representation-Mediated

(1) Formulation

Representation-Mediated Modeling offers a new pathway that connects the understanding module with the generation module through intermediate representations. The key idea behind this modeling approach is that an LLM first processes the input to generate intermediate representations, which then control the downstream generative models. These representations are then used as conditional inputs by external generative models to complete the target modality generation task. Compared to Prompt-Mediated Modeling, this approach discards natural language as an intermediary expression form and focuses more on the structural alignment of the model's internal semantic space, providing clear advantages in terms of information fidelity and generation control precision.

Representation-Mediated Modeling features a loose coupling structure between understanding and generation. Although the generation module and the understanding module remain relatively independent in terms of training and structure, they can be connected through an intermediate aligned semantic space. This "mediated alignment" not only enhances generation quality in multimodal tasks but also endows the model with stronger structural expression capabilities and fine-grained control. Particularly for generation tasks involving complex image structures or interwoven multimodal content, the high-density semantic representations offered by Representation-Mediated Modeling can effectively mitigate the information loss issues that arise when using natural language prompts for modality translation.

In Representation-Mediated Modeling, the model architecture is typically based on an autoregressive framework to fully leverage the strong representation and generation capabilities of existing LLMs. As for the generation module, external generators usually use diffusion-based models due to their excellent generation quality in continuous modalities such as images and audio. However, since the output representations of LLMs are often in embedding spaces that are inconsistent with the input space of diffusion models, directly using them as conditional inputs to the generative model is not advisable. To address this issue, existing research has proposed two mainstream solutions: one approach [137] involves training the generative model itself to adapt to the output representations of the LLM, while the other [194], [131], [132], [139], [140] uses a more lightweight solution by designing specialized intermediate representation transformation modules to map the LLM output into semantically meaningful regions of the generative model's input space.

(2) Relevant Works

Throughout the development of Representation-Mediated Modeling, researchers have extensively explored key aspects such as the formulation of intermediate representations, structural connections, and optimization strategies, achieving a series of notable breakthroughs.

Continuous Input. The continuous or discrete nature of multimodal inputs has a substantial impact on how the model processes and integrates information. To mitigate the information loss often caused by operations like Vector Quantization (VQ), most existing approaches adopt continuous input features—for example, GILL [194], VL-GPT [195], MiniGPT-5 [134], MM-Interleaved [139], BLIP3-o [196], OmniGen2 [176], and UniLIP [197].

GILL [194] is the first to demonstrate that the language space of LLMs can be aligned with the embedding space of image generation models via feature mapping, enabling effective use of pretrained generators. It further introduces GILLMapper, which maps the `[IMG]` token embedding into the embedding space of Stable Diffusion for image generation. The Emu series [129], [137] further advances the practice of joint modeling for text-image sequences. Emu[129] adopts a unified autoregressive objective over interleaved text, image, and video data, classifying text tokens and regressing visual embeddings. Images are first encoded as

2D spatial embeddings using EVA-CLIP[53], then converted into 1D sequences via a causal transformer before being fed into the LLM. Emu2 [137] simplifies this pipeline by removing the causal transformer from the encoder. During training, it directly decodes visual embeddings into images, effectively decoupling image generation from the LLM.

VisionLLM v2 [198] significantly expands the application scope by integrating over a hundred visual tasks within a unified framework. It introduces a super link mechanism, which connects the MLLM to task-specific decoders via a routing token, enabling flexible task information transfer and gradient feedback. In contrast, TextHarmony [146] focuses on text-centric tasks, particularly the understanding and generation of visual text. It proposes the Slide-LoRA architecture, which dynamically aggregates modality-specific and modality-agnostic LoRA experts. This partially decouples the multimodal generation space, thereby alleviating conflicts between generation tasks across different modalities. Beyond task specialization, recent work has explored fine-grained control and diversity in visual generation. PUMA [151] improves performance on tasks like image editing and inpainting by unifying multi-granularity visual features as both input and output to the MLLM. It uses a CLIP-based encoder to extract features at different semantic levels and fine-tunes diffusion-based decoders accordingly. These decoders effectively recover missing details from coarse features and preserve precision for fine-grained ones, making them well-suited for multi-scale generation. WeGen [199] targets subject-driven generation by addressing instance identity consistency. It proposes a "Dynamic Instance Identity Consistency" pipeline to balance fixed and variable attributes, and enhances diversity through Prompt Self-Rewriting, which introduces randomness via discrete token sampling while maintaining semantic alignment.

Training strategies have emerged as a key enabler of Representation-Mediated Modeling. MetaMorph[155] introduces visual-predictive instruction tuning, a lightweight approach that converts an LLM into a UFM with minimal changes. It connects a vision head to Stable Diffusion [15] and trains the model using cosine similarity between predicted visual tokens and those from a vision encoder. Joint training shows that visual understanding and generation can reinforce each other, with understanding data boosting both capabilities. Nexus-Gen[168] tackles the mismatch between training and inference in autoregressive models, which leads to error accumulation in continuous image embeddings. It proposes a prefilling strategy that uses special tokens as placeholders during both the training and inference phases, ensuring consistency.

In parallel with training strategies, several studies have focused on visual representation design to improve unified modeling. BLIP-3o [23] conducts a comprehensive investigation into image encoding, architectural choices, and training approaches. Its experiments yield three key findings: (1) CLIP-based representations outperform VAE-based ones in both training efficiency and generation quality; (2) using flow matching loss instead of MSE improves generation diversity; and (3) a two-stage training strategy—first training an autoregressive model for vision understanding, then freezing it to train the generation model—outperforms multitask joint training. These insights offer valuable guidance for the design of UFMs. UniWorld-V1 [200], inspired by observations from GPT-4o [10], extracts visual features via a semantic encoder and achieves effective integration of image understanding, generation, and editing with only 2.7M training samples. Pisces [201] adopts a representation-decoupling strategy similar to Janus [150], using EVA-CLIP [53] and SigLIP [33] for separate visual embeddings.

Discrete Input. In contrast to continuous inputs, discrete input features align more naturally with the autoregressive nature of LLMs, enabling unified modeling through next-token prediction in a tokenized space. However, since mainstream visual encoders such as CLIP-ViT [31] produce continuous representations, many approaches adopting discrete inputs redesign or enhance the tokenizer component. Notable examples include the SEED [130] and LaVIT [133], which introduce improved visual tokenization schemes tailored for alignment with language modeling. MIO [202] also directly utilizes the SEED tokenizer [130] to convert images into discrete tokens, facilitating seamless integration with the LLM, and further demonstrates the effectiveness of this design through empirical validation.

Tokenizer Innovation. As noted above, to better integrate visual modalities into the autoregressive framework of LLMs, many works have put forward a variety of innovations in visual tokenization.

SEED-OPT [130] introduces the SEED tokenizer, which encodes images into a sequence of discrete image tokens for autoregressive modeling. A key insight is that these tokens should be represented as 1D sequences reflecting causal dependency, independent of their original 2D patch positions. Building on this, SEED-LLaMA [135] further simplifies the architecture by replacing the original reverse Q-Former with a Multilayer Perceptron (MLP). SEED-X [130] further extends the framework by introducing dynamic resolution and multi-granularity de-tokenization, enabling the model to understand images of arbitrary sizes and aspect ratios, while also supporting image generation at varying levels of granularity. Similar to SEED [130], LaVIT [133] also designs a dedicated tokenizer to convert images into discrete token sequences—treating visual inputs as a "foreign language" understandable by LLMs. To account for varying semantic complexity across images, LaVIT adopts a hybrid strategy combining token selection and token merging to produce discrete visual token sequences with dynamic lengths. Video-LaVIT [141] extends this approach to the video domain by decomposing videos into static keyframes and dynamic temporal motions, which are encoded separately, enabling effective handling of video tasks.

ILLUME [164] introduces a semantic visual tokenizer that enables image-text alignment pretraining using only 15M data samples. Unlike traditional VQ-based tokenizers trained with reconstruction loss, the semantic tokenizer discretizes images into tokens in a semantic feature space. Building on this, ILLUME+ [203] proposes DualViTok, a dual-branch vision tokenizer that separately captures high-level semantic concepts and low-level texture details. During image generation, it adopts a coarse-to-fine image representation strategy—first generating semantic tokens, then texture tokens—thereby accelerating the alignment between text and fine-grained visual details and improving both high-fidelity and semantic consistency in the generated im-

ages. DDT-LLaMA [204] identifies a common limitation in relying solely on spatial visual tokens extracted from image patches. To overcome this, it introduces Discrete Diffusion Timestep (DDT) tokenization, which enables image reconstruction at arbitrary diffusion steps.

Query-Based Composition. Some works adopt query-based composition to bridge the LLM and the generative model. In these approaches, a set of learnable queries serves as an interface that efficiently aggregates condition features required for multimodal generation. This design provides a flexible and effective mechanism to connect the understanding and generation modules within UFMs.

DreamLLM [132] introduces dream queries to aggregate prior information as conditions for image generation, serving as a lightweight interface without altering the LLM's output space. Additionally, instead of relying on visual encoders like CLIP [31], DreamLLM performs direct sampling in the original multimodal space, effectively avoiding the information loss typically introduced by intermediate encoders. MetaQuery [166] introduces a set of "MetaQueries" as the interface between a frozen MLLM and a diffusion model, aiming to effectively transfer the MLLM's understanding and reasoning capabilities into the image generation process. This design avoids the complexity and imbalance that may arise from having the MLLM generate tokens across different modalities. Specifically, the MetaQueries are fed into the MLLM to query out generation conditions, which are then passed to the diffusion model through a lightweight connector. By leveraging MetaQueries, researchers can transfer existing models without building end-to-end architectures from scratch, requiring only standard image-text data and diffusion objectives for training. For example, Ming-Lite-Uni [169] integrates this mechanism with M2-Omni [192], introducing multi-scale learnable tokens and a multi-scale representation alignment to enhance cross-modal understanding and generation. The follow-up work, Ming-Omni [174], further adopts an MoE architecture and incorporates audio modalities, significantly expanding its multimodal capabilities and achieving modality support comparable to that of GPT-4o [10]. OpenUni [173] builds on MetaQuery [166] with a more lightweight implementation—for example, reducing the connector module to just 6 transformer layers. Despite using fewer parameters, it still achieves competitive performance, offering valuable insights into the simplified design of UFMs.

Modality Integration. Similar to the Prompt-Mediated Modeling strategy, Representation-Mediated Modeling benefits from its modular and decoupled structure, making it relatively easy to extend beyond text and image to other modalities such as video and audio.

For video tasks, NExT-GPT [131] introduces modality-specific adaptors at both the input and output stages to align features with the text space and target modality space. This design allows the model to leverage existing pretrained encoders and decoders for any-to-any multimodal generation. Since only a few projection layers within the adaptors need to be finetuned, NExT-GPT achieves efficient semantic alignment while significantly reducing training costs. Recognizing that cross-modal alignment is crucial in building UFMs, X-VILA [205] addresses the problem of severe visual information loss during modality extension. To mitigate this in video generation, it proposes a visual embedding highway mechanism that bypasses the LLM by directly feeding encoder outputs—processed via zero-convolution—into the generative model as conditional input. This approach provides more direct guidance and helps preserve fine-grained visual details during generation.

Following the progress in the video domain, Representation-Mediated Modeling has also begun to extend into the audio modality. CoDi-2 [136], as an improved version of CoDi [128], incorporates an MLLM as the fundamental engine, enhancing the model's ability to understand and reason over complex instructions. However, AnyGPT [140] points out a key limitation in models like NExT-GPT [131] and CoDi-2 [128]: the use of separately trained encoders and decoders may lead to inconsistencies between the LLM's input and output representations. To address this, AnyGPT encodes both image and audio modalities into discrete tokens and feeds them directly into the LLM, enabling the model to semantically perceive and process multimodal inputs in a unified manner. To further improve audio generation—especially in open-domain settings—C3LLM [206] adopts a hierarchical generation strategy built on an LLM backbone. Specifically, the LLM first generates coarse-grained acoustic tokens, which are then refined into fine-grained ones using a non-autoregressive transformer, ultimately producing high-fidelity audio. Meanwhile, MuMu-LLaMA [157] and LM-MSN [207] explore other dimensions of improvement. Specifically, MuMu-LLaMA emphasizes the construction of high-quality training datasets, while LM-MSN focuses on the design of a variational quantization mechanism.

(3) Pros and Cons

The outstanding advantage of Representation-Mediated Modeling lies in its high density and structurality in information representation. Intermediate features can transmit semantic details more precisely than natural language, thus demonstrating higher generation quality and control capabilities. Representation-Mediated Modeling can also fully leverage the leading image generation capabilities of diffusion models, producing high-quality images. Additionally, this paradigm supports joint training optimization with external diffusion models, which can further enhance the model's generation consistency and overall performance.

However, this modeling approach also faces certain challenges. On the one hand, the additional computational overhead and structural complexity brought by joint training require higher engineering investment. On the other hand, compared to Prompt-Mediated Modeling, Representation-Mediated Modeling requires alignment of intermediate features with the input space of the diffusion model. Poor alignment can significantly impact generation quality. Furthermore, since the model's generation capability is still constrained by the external generative module itself, its scalability may face bottlenecks as the model size increases.

Therefore, while Representation-Mediated Modeling offers notable advantages in generation quality and controllability, it remains a compromise solution for achieving unified generative capabilities. To enable deeper integration between multimodal understanding and generation, further advances are needed—particularly through more tightly

coupled architectural designs.

3.3 End-to-End Unified Modeling

Compared to the previous two modular modeling approaches, End-to-End Unified Modeling represents a more tightly integrated paradigm. This approach emphasizes jointly modeling multimodal understanding and generation within a unified architecture through end-to-end training. Unlike modular methods that rely on external generation modules, End-to-End Unified Modeling performs both perception of input modalities and generation of target modalities entirely within the model itself. This results in a higher degree of coupling and semantic consistency. On the one hand, the tight integration helps reduce information loss during modality transformation. On the other hand, it enables the model to handle multimodal tasks within a shared representation space, thereby enhancing both its expressive capacity and generation quality.

Fig. 6 illustrates the taxonomy of End-to-End Unified Modeling. Based on differences in model architectures, generation mechanisms, and their underlying principles, current End-to-End Unified Modeling approaches can be roughly categorized into the following four types: (1) **Autoregressive Modeling** (Sec. 3.3.1), (2) **Diffusion Modeling** (Sec. 3.3.2), (3) **Autoregressive-Diffusion Hybrid Modeling** (Sec. 3.3.3), and (4) **Other types** (Sec. 3.3.4).

3.3.1 Autoregressive

(1) Definition

Among the various implementation paths for End-to-End Unified Modeling, Autoregressive Modeling is the most common and mature approach. The basic idea is to encode input information from different modalities into the token sequence, and then generate the output sequence step by step through an autoregressive mechanism. These models typically use decoder-only architectures, combined with the causal masking strategy, ensuring that the generation of each token depends only on previously generated content, thereby maintaining the orderliness and structure of the generation process. During the training phase, the model simultaneously optimizes both understanding and generation objectives through multitask learning, endowing it with natural context modeling capabilities. This enables flexible handling of tasks such as image-to-text, text-to-image, and even cross-modal interleaving tasks.

This modeling paradigm typically constructs a unified input sequence by concatenating multimodal tokens generated by the modality-specific tokenizer with text tokens, and extends the vocabulary to support the generation of multimodal tokens such as images and audio. The training objective is to maximize the conditional probability of each token in the sequence, thereby achieving end-to-end multimodal joint modeling. Since there is no need to introduce separate generative modules, this approach is easier to implement for system-level joint optimization, offering strong parameter-sharing capabilities and scalability.

(2) Relevant Works

Discrete Input. To better align with the characteristics of next token prediction, most existing works apply operations such as vector quantization to multimodal feature inputs, converting them into discrete tokens.

Emu3 [25] is one of the most representative works in this line of research. It unifies text, image, and video inputs into discrete tokens through tokenization and employs a single transformer architecture to train from scratch. Relying solely on next-token prediction, Emu3 demonstrates strong performance across a wide range of multimodal tasks, showcasing the great potential of this paradigm. Compared to its predecessors Emu [129] and Emu2 [137], Emu3 abandons diffusion-based generation in favor of a more streamlined architecture. To further improve visual generation quality, it introduces Quality Fine-Tuning (QFT) and Direct Preference Optimization (DPO), which finetune the model using high-quality multimodal data and human preference signals, respectively. Building upon this foundation, Emu3.5 [210] further validates the feasibility of the next-token prediction paradigm. It not only exhibits strong native multimodal capabilities but also begins to demonstrate generalizable world-modeling abilities, suggesting a deeper integration of perception, reasoning, and generation within a unified framework. LWM [142] leverages architectural optimizations such as RingAttention to scale context length up to 1M tokens, enabling the training of a language model capable of handling extremely long text sequences. Building upon this foundation, LWM introduces visual modalities into training, allowing the model to perform both image and video understanding and generation. During training, the order of textual and visual data is randomly shuffled to support diverse tasks such as image understanding and text-to-image generation. To balance the distinct characteristics of different modalities, LWM also incorporates an efficient masked sequence packing strategy. Liquid [156] follows a similar paradigm by directly feeding discretized image tokens obtained via VQGAN [96] into the LLM. Extensive experiments based on LLaMA 3 [236], Gemma 2 [237], and Qwen2.5 [238] reveal that training on multimodal data can initially degrade language performance compared to text-only training. However, this gap diminishes as model size increases, suggesting that larger models are more capable of jointly handling both understanding and generation. Moreover, increasing training data for either understanding or generation leads to performance gains in the other, highlighting the mutual benefits of this unified modeling approach. To address the complexity of training UFMs, UGen [208] adopts a strategy called progressive vocabulary learning, which incrementally activates visual token IDs during training. These tokens are gradually integrated into the training process, enabling the model to progressively acquire image understanding and generation capabilities. SpeechGPT [121] and ARMOR [209] extend the exploration of unified modeling into the audio domain and resource-constrained settings, respectively.

Hybrid Input. Considering that visual understanding tasks typically require high-level semantic information, while visual generation tasks rely more on low-level structural and textural details, using a single type of input, whether discrete or continuous, may lead to a trade-off between the two tasks. To mitigate this conflict caused by the differing granularity of required information, some works [150], [26], [211], [165] by separately extracting both

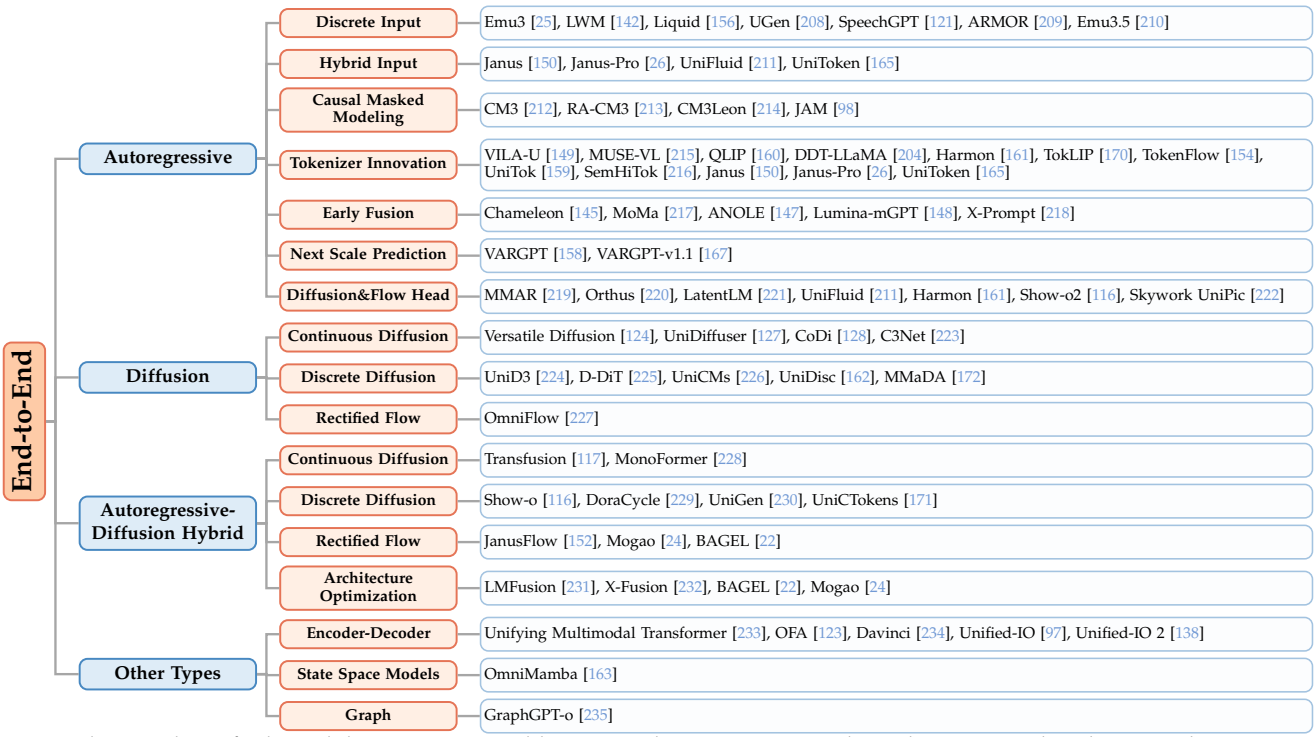

Fig. 6: End-to-End Unified Modeling, categorized by input characteristics, technical routes, and architectural innovations.

discrete and continuous visual features.

Among them, Janus [150] and Janus-Pro [178] utilize SigLIP [33] and a VQ tokenizer [239] to extract high-level semantic features and low-level visual details, respectively, which are then used for understanding and generation tasks. This decoupled design improves model flexibility and has been empirically shown to alleviate the conflicts and performance bottlenecks associated with using a single visual encoder. Moreover, such a modular setup enables extensibility—for instance, the SigLIP encoder can be replaced with a more powerful visual backbone—allowing Janus to potentially scale to additional modalities such as video. UniFluid [211] follows a similar design philosophy by leveraging different visual representations for different tasks. However, using task-specific visual encoders may limit the overall flexibility of the unified model. To address this, UniToken [165] proposes concatenating both continuous and discrete visual features into a unified visual representation, thereby providing comprehensive visual input.

Causal Masked Modeling. CM3 [212] is one of the earliest representative works to achieve unified processing of multimodal tasks using an autoregressive approach. The key innovation lies in the introduction of the "Causal Masked Modeling" method, which combines traditional causal language modeling with masked language modeling. By moving the masked portion to the end of the sequence, the model simulates the masking effect during autoregressive generation, thereby introducing bidirectional context information. This approach was later inherited and extended by subsequent works. RA-CM3 [213], built on the foundation of CM3, introduces a retrieval-augmented mechanism that guides the model to fetch relevant text or images from an external multimodal document library, reducing dependence on the model's parameter size. Furthermore, CM3Leon [212] simplifies and extends RA-CM3 [213] by adopting a two-stage training strategy, which includes large-scale retrieval-augmented pretraining and multitask supervised fine-tuning. JAM [98] proposes a modular fusion framework based on CM3Leon, which connects existing language models and image generation models through cross-modal attention, resulting in a unified autoregressive model capable of generating both text and images. Notably, JAM achieves this with less than 1% of the training data required by the original model.

Tokenizer Innovation. Traditional VQ-based tokenizers, such as VQGAN [96], are typically trained on image reconstruction tasks and thus tend to discard a significant amount of semantic information. This limitation can substantially impair the understanding capabilities of UFMs. As a result, improving and innovating tokenizer design has become a critical and promising direction in the context of Autoregressive Modeling for UFMs.

To better align with the textual modality, VILA-U [149] introduces a unified vision tower, which integrates both reconstruction loss and contrastive loss into the training of its autoencoder. Furthermore, residual quantization is utilized to enrich the visual representations, allowing the model to capture high-level semantics comparable to CLIP [31] while still retaining the ability to perform image reconstruction in the style of VQGAN [96]. However, VILA-U [149] suffers from loss conflicts during training. To address this issue, MUSE-VL [215] proposes Semantic-aware Discrete Encoding (SDE) by incorporating semantic features from a CLIP-style model. These features are fused with visual representations from the encoder and subsequently used in the quantization process. To balance reconstruction and semantic preservation, MUSE-VL employs a dual-branch decoder design. QLIP [160] introduces a binary-spherical

quantization-based autoencoder, which effectively balances image reconstruction and semantic understanding through a two-stage training scheme. In contrast, TokLIP [170] proposes a different strategy by avoiding the discretization of CLIP features. Instead, it semanticizes VQ tokens, adopting a discrete-to-continuous approach that decouples the training objectives of understanding and generation. This design simplifies training while leveraging CLIP-level semantic representations more effectively.

Although most existing methods rely on spatial visual tokens extracted from image patches and arranged in a spatial order, DDT-LLaMA [204] argues that such tokens lack the hierarchical structure inherent to natural language, making them suboptimal as a linguistic interface. To overcome this limitation, DDT-LLaMA designs recursive visual tokens learned through diffusion timestamps, which can progressively recover attributes lost in increasingly noisy images. By incorporating a diffusion timestep tokenizer, DDT-LLaMA enables reconstruction of the original image from any timestep, thus bridging diffusion-based generation with the LLM effectively. Given the distinct granularity of visual information required for understanding and generation tasks, relying on a single VQ-based encoder often leads to performance trade-offs. To address this challenge, several approaches explore multi-codebook or hierarchical codebook designs, as seen in TokenFlow [154], UniTok [159], and SemHiTok [216], enabling more flexible and task-specific visual representations. Other methods adopt a dual-encoder architecture to explicitly separate visual processing for different tasks—for example, the aforementioned Janus [150], Janus-Pro [26], and UniFluid [211].

Early Fusion. A line of work led by Chameleon [145] adopts an early fusion strategy, where textual and non-textual modalities are projected into a shared representation space from the very beginning. This approach encourages deep cross-modal integration and facilitates more effective multimodal interaction. The modeling approach of this line of work is similar to that of works such as Emu3 [25]. However, given their extensive experimental analyses and the large number of subsequent extensions they have inspired, we organize these works separately.

Chameleon [145] systematically investigates unified autoregressive modeling for multimodal understanding and generation, providing key insights into architectural design and training stabilization. It employs a unified transformer backbone trained end-to-end, enabling deep modality fusion and high performance, but at the cost of increased training instability and computational overhead as model size grows. To address these challenges, Chameleon introduces architectural and optimization enhancements, such as query-key normalization and revised layer normalization placement, which improve convergence and robustness. To further enhance pretraining efficiency, MoMa [217] builds on Chameleon's early-fusion strategy by introducing a modality-aware MoE architecture, grouping experts by modality to improve computational efficiency while preserving cross-modal integration. MoMa also explores sparse scaling along both depth and width, demonstrating accelerated convergence of pretraining objectives. Subsequent works, including ANOLE [147], Lumina-GPT [148], and X-Prompt [218], extend the Chameleon framework with targeted improvements. For example, Lumina-GPT proposes Flexible Progressive Supervised Finetuning (FP-SFT), a weak-to-strong curriculum that gradually adapts the model from low- to high-resolution image tokens. X-Prompt introduces a method for compressing in-context signals into fixed-length token representations, alleviating context window limitations in in-context learning.

Next Scale Prediction. Inspired by the concept of VAR [115], VARGPT [158] further integrates next token prediction (for visual understanding) and next scale prediction (for visual generation), exploring a new paradigm of unified autoregressive generation and expanding the boundaries of unified modeling. The improved version, VARGPT-v1.1 [167], introduces enhancements in training strategy, data quality, and model backbone. It emphasizes progressively increasing image resolution while iteratively applying instruction tuning and reinforcement learning, aiming to further boost the model's capabilities in both visual understanding and generation.

Diffusion&Flow Head. Considering that vector quantization and related approaches often incur substantial information loss, some studies [219], [220], [221], [211], [161], inspired by works such as MAR [240], adopt an autoregressive backbone while introducing a lightweight diffusion head or flow head. This design aims to retain as much information as possible during the generation process, thereby mitigating the degradation caused by aggressive quantization.

MMAR [219] was among the first to empirically show that hybrid approaches such as Transfusion [117] and MonoFormer [228], which combine diffusion and autoregressive models within a shared backbone, require noise injection to enrich semantic representations for image generation. However, this noise also induces significant information loss, impairing understanding performance. To address this, MMAR, inspired by MAR, decouples the diffusion process from the autoregressive backbone by introducing a lightweight MLP-based diffusion sampler. This sampler leverages outputs from the autoregressive transformer to sample continuous image tokens, enabling image generation without the detrimental effects of noise injection. LatentLM [221] encodes continuous data into latent vectors via a VAE, concatenates them with text tokens, and feeds the sequence into a causal transformer. The predicted visual tokens are then processed by a diffusion head and decoded into images using the VAE decoder. To prevent variance collapse, LatentLM adopts a fixed-variance "σ-VAE" in the latent space. Similarly, UniFluid [211] employs a dual-vision encoder to decouple representations for understanding and generation, utilizing a per-token diffusion head for image synthesis. UniFluid demonstrates that, with carefully tuned training strategies, understanding and generation tasks can be mutually beneficial, effectively balancing their respective losses. Moreover, the choice of LLM backbone substantially influences image generation quality. Orthus [220] and Harmon [161] also adopt diffusion heads for image generation, while Skywork UniPic [222] follows a similar structure to Harmon but incorporates SigLIP2 for understanding. In contrast, Show-o2 [175] employs a lightweight flow matching head as an alternative for image generation.

(3) Pros and Cons

In summary, Autoregressive Modeling has rapidly advanced in architecture, representation, and training efficiency, establishing a robust foundation for high-quality, consistent UFMs. This paradigm aligns naturally with mainstream LLMs, requiring no complex auxiliary modules and maintaining architectural simplicity and ease of implementation. The unified decoder structure accommodates both language and non-language sequences, offering strong scalability. Its end-to-end autoregressive mechanism enables joint optimization of understanding and generation within a shared semantic space, facilitating parameter sharing and cross-task knowledge transfer.

However, several challenges remain. First, due to substantial differences across modalities, autoregressive models may struggle to align multimodal features effectively without access to sufficiently large and high-quality datasets, leading to high training costs. Second, reliance on token-based reconstruction for image generation introduces information bottlenecks, limiting image fidelity and trailing diffusion-based methods. Third, error accumulation during autoregressive generation, especially for long sequences, can degrade performance. Finally, the inherently sequential generation process leads to low inference efficiency, posing difficulties for real-time applications.

3.3.2 Diffusion

(1) Definition

Diffusion models have achieved remarkable success in the field of image generation and have gradually become one of the mainstream generative paradigms, thanks to their advantages in generation quality, diversity, and training stability. Inspired by this success, researchers have begun exploring the extension of the diffusion modeling paradigm to a broader range of multimodal generation tasks, with UFMs being one of the key application directions.

We refer to the modeling paradigm based on diffusion as "Diffusion Modeling". Based on the underlying principles and technical characteristics of diffusion modeling, we further divide it into three representative categories: continuous diffusion, discrete diffusion, and rectified flow. In the following sections, we discuss each category in detail through representative works.

(2) Relevant Works

Recently, the scalability, unification, and controllability of diffusion models in multimodal modeling have attracted widespread attention, motivating researchers to propose various methods to explore and enhance these capabilities.

Continuous Diffusion. Versatile Diffusion [124] is one of the earliest diffusion-based UFMs to support multimodal generation. Built upon a single-flow diffusion framework, it extends to a multi-flow diffusion mechanism, where each task flow is regarded as generating features of modality n conditioned on modality m. By constructing multiple task flows, sharing the generative module, and partitioning the context encoder, the model improves parameter efficiency and scalability while demonstrating strong performance on tasks such as bidirectional text-image generation and multi-style image fusion. Going further, UniDiffuser [127] aims to build a UFM that simultaneously captures the marginal, conditional, and joint distributions of multimodal data. It formulates all generation tasks as a noise prediction problem—despite the different perturbation schemes across modalities, the objective is unified as modeling the noise in corrupted data. Building on this foundation, CoDi [128] realizes any-to-any multimodal generation by training a latent diffusion model for each modality, incorporating a cross-attention module into the diffuser, and projecting features from all modalities into a shared latent space, thereby enabling conditional generation across arbitrary modality combinations. Given the scarcity of real paired multimodal training data, CoDi adopts input-output space alignment instead of full joint distribution modeling, significantly improving sample efficiency. However, its use of linear interpolation in latent space may degrade generation quality. To mitigate this, C3Net [223] proposes an improvement: it constructs a unified Control C3-UNet module based on ControlNet [17], which aligns each modality into a shared semantic latent space, then performs generation from the fused representation. This avoids the quality degradation caused by latent space interpolation.

Discrete Diffusion. In addition to conventional continuous diffusion models, discrete diffusion [241], [242] has garnered increasing attention from researchers due to its outstanding advantages in generation quality.

UniD3 [224] is one of the first to introduce discrete diffusion into a unified modeling framework. It enhances text-image fusion and content consistency through a mutual attention module with fused embeddings, and designs a transition matrix to achieve implicit cross-modal alignment. D-DiT [225], building on the Multimodal Diffusion Transformer (MM-DiT) [103], adopts a dual-branch architecture that combines continuous diffusion for images with discrete diffusion for text, leveraging cross-modal attention to strengthen modality interactions. UniCMs [226] introduces a novel consistency distillation method, unifying text and image generation as discrete denoising trajectories. It also proposes a trajectory segmentation strategy that improves the stability and convergence of multimodal generation via distillation. UniDisc [162] constructs a fully discrete diffusion-based unified structure, jointly tokenizing both images and text, and uses self-attention mechanisms to reconstruct masked sequences. This approach significantly boosts performance in both multimodal generation and discrimination, outperforming autoregressive methods across metrics such as FID and CLIP Score. Additionally, MMaDA [172] also builds on discrete diffusion, and further optimizes the post-training phase by introducing mixed long chain-of-thought finetuning and UniGRPO, enhancing the model's ability to handle complex tasks.

Rectified Flow. Since rectified flow can essentially be regarded as an improved variant of diffusion modeling, we categorize and analyze rectified flow-based approaches under this paradigm as well. OmniFlow [227] is built upon the rectified flow framework and extends it to multimodal joint modeling. It adopts a modular design that eliminates the need for training from scratch. By introducing additional input-output flows on top of MM-DiT, OmniFlow expands its capabilities from text-to-image generation to general any-to-any generation. Owing to the largely independent parameterization of different modality-specific flows, each flow can be pre-trained separately or initialized using existing

single-task expert models.

(3) Pros and Cons

Diffusion Modeling is characterized by its high-quality and detail-rich generation capabilities. However, it also exhibits notable limitations, including slow inference speed due to the multi-step denoising process and relatively weak multimodal understanding. Despite these challenges, recent diffusion-based approaches have made significant strides in enhancing both unification and generation quality. With the rapid development of Diffusion Language Models (DLMs) such as LLaDA [243], [244], diffusion modeling is expected to hold even greater research potential in the field of UFMs. Collectively, they reflect a clear trend toward more generalizable and flexible unified modeling, capable of supporting bidirectional and even arbitrary modality conversions.

3.3.3 Autoregressive-Diffusion Hybrid

(1) Definition

Autoregressive and diffusion models represent the two most prominent modeling paradigms in current multimodal generation tasks, each exhibiting distinct advantages in handling discrete modalities (e.g., text) and continuous modalities (e.g., images, audio), respectively. Autoregressive models have long dominated the field of text generation due to their strong sequential modeling capabilities and high-quality language outputs. In contrast, diffusion models excel in tasks such as image generation and image editing, thanks to their powerful ability to model high-dimensional continuous spaces.

However, as UFMs continue to evolve, a persistent challenge remains: how to effectively handle both discrete and continuous modalities within a single framework, while achieving strong performance in both understanding and generation. Given the fundamental differences between modalities like text and images—in terms of data structure, semantic representation, and modeling requirements—some recent works have begun exploring tighter integration of these two modeling paradigms within a unified architecture, aiming to leverage the strengths of both to achieve more effective understanding and generation. We refer to this type of modeling approach as "Autoregressive-Diffusion Hybrid Modeling". In Autoregressive-Diffusion Hybrid Modeling, the model is required to learn both language modeling and diffusion modeling simultaneously. Given the close relationship between Rectified Flow [245] and traditional diffusion models, we also include approaches that integrate autoregressive modeling with rectified flow under this category for unified analysis.

(2) Relevant Works

Continuous Diffusion. Transfusion [117] stands out as one of the earliest representative works to implement Autoregressive-Diffusion Hybrid Modeling. This method employs a single transformer architecture with two distinct training objectives: a next-token prediction loss for text, and a diffusion loss for images, defined respectively as:

$$\mathcal{L}_{\text{LM}} = \mathbb{E}_{y_i}\left[-\log P_\theta(y_i \mid y_{<i})\right], \tag{20}$$

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)\right\|^2\right]. \tag{21}$$

This dual-objective setup enables efficient and unified handling of both discrete and continuous modalities under the shared data and parameter settings. To better jointly optimize the two different objectives, Transfusion combines them through a weighted summation over modalities using a balancing coefficient λ:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}. \tag{22}$$

Through ablation studies, Transfusion finds that incorporating a bidirectional attention mechanism for image modeling significantly outperforms traditional causal attention, as it allows each image patch to perceive the entire image—preserving structural integrity and continuity. After scaling the model to 7 billion parameters and training on over 2 trillion multimodal tokens, Transfusion achieves performance on par with pure diffusion models or language models of the same scale, demonstrating the feasibility and scalability of this hybrid modeling paradigm. MonoFormer [228] adopts a similar architecture to Transfusion by combining continuous diffusion modeling with autoregressive mechanisms, and also uses bidirectional attention for visual modeling. What distinguishes MonoFormer is its experimental verification that initializing the transformer with a well-pretrained LLM could significantly improve training efficiency and overall performance.

Discrete Diffusion. Developed contemporaneously with Transfusion [117], Show-o [116] adopts a similar Autoregressive-Diffusion Hybrid Modeling framework, but specifically implements it using discrete diffusion. Specifically, Show-o adopts two learning objectives: Next Token Prediction (NTP) for autoregressive modeling and Masked Token Prediction (MTP) for discrete diffusion modeling. These objectives are defined as follows:

$$\mathcal{L}_{\text{NTP}} = \sum_i \log p_\theta\left(y_i \mid y_1, \ldots, y_{i-1}, y_1, \ldots, y_N\right), \tag{23}$$

$$\mathcal{L}_{\text{MTP}} = \sum_j \log p_\theta\left(x_j \mid x_*, x_2, \ldots, x_*, x_N, x_1, \ldots, x_M\right), \tag{24}$$

where x and y denote image tokens and text tokens, respectively, while M and N represent their corresponding sequence lengths. x_* refers to image tokens in the input sequence that have been replaced by `[MASK]` tokens. The overall loss is a weighted combination of the two:

$$\mathcal{L} = \mathcal{L}_{\text{MTP}} + \lambda \cdot \mathcal{L}_{\text{NTP}}, \tag{25}$$

where λ is a balancing coefficient controlling the contribution of the autoregressive objective. Show-o adopts Omni-Attention to enable bidirectional interaction among image tokens and unidirectional modeling for text tokens, mirroring Transfusion's attention strategy. To further unify the understanding and generation processes across modalities, Show-o introduces a unified prompting technique that uses special tokens such as `[MMU]` and `[T2I]` to explicitly specify the different task types.

The strong performance of Show-o has inspired many subsequent works [229], [230], [171] to build upon and improve it. For example, DoraCycle [229], built on top of Show-o, proposes a dual-modality cyclic mechanism (text $\rightarrow$ image $\rightarrow$ text and image $\rightarrow$ text $\rightarrow$ image) to achieve domain adaptation without paired data. It fully leverages the bidirectional vision-language mapping ability acquired during the pretraining stage, showcasing the potential for

self-evolution in UFMs. Additionally, UniCTokens [226] introduces a set of concept tokens to enhance the model's capacity for understanding and generating semantic concepts. Through a three-stage training process, it enables information complementarity between understanding and generation tasks, further advancing the performance.

Rectified Flow. In addition to integration with standard diffusion modeling, some studies choose to combine rectified flow with autoregressive modeling to enhance generation quality. JanusFlow [152] builds upon the decoupled encoder design from Janus [150], preserving both high-level semantics and low-level visual details more effectively. It seamlessly integrates rectified flow with the LLM architecture and enables flow operations using a lightweight encoder-decoder setup, significantly simplifying the overall system design. Recent works such as BAGEL [22] and Mogao [24] also adopt rectified flow. BAGEL [22] emphasizes leveraging high-quality, large-scale interleaved data to strengthen multimodal understanding and generation. Experimental results show that as the scale of training data increases, the model not only excels in standard tasks but also exhibits emergent capabilities in long-context visual reasoning, substantially extending the application boundaries of UFMs. Mogao [24] also advances the exploration of interleaved multimodal modeling by introducing a novel multi-conditioned image generation paradigm. Specifically, it dynamically integrates contextual information from both preceding text and visual content, enabling more coherent and context-aware generation.

Architecture Optimization. To address task interference issues observed in models such as Show-o [116] and Transfusion [117], which share transformer backbone parameters, and to improve both training efficiency and overall model performance, researchers have proposed a series of architectural innovations, such as the introduction of MoE.

LMFusion [231] builds upon the hybrid modeling strategy of Transfusion [117] by introducing a modality-aware MoE architecture. This design facilitates interaction between different modalities within a unified self-attention framework. To accommodate modality-specific characteristics, the model employs distinct QKV projections and feed-forward networks (FFNs) for textual and visual inputs, enabling separate yet coordinated processing across modalities. A similar design is also employed by BAGEL [22], Mogao [24], and MoMa [217]. The framework can be expressed as:

$$(Q_i, K_i, V_i) = \begin{cases} (x_i W_{Q_V}, x_i W_{K_V}, x_i W_{V_V}), & \text{if } x_i \text{ is visual} \\ (x_i W_{Q_T}, x_i W_{K_T}, x_i W_{V_T}), & \text{if } x_i \text{ is textual} \end{cases}, \tag{26}$$

$$\begin{aligned} \text{Attention Output} &= \text{SharedSelfAttention}(Q, K, V) \\ &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \end{aligned}, \tag{27}$$

$$\text{FFN Output}_i = \begin{cases} \text{FFN}_V(z_i), & \text{if } x_i \text{ is visual} \\ \text{FFN}_T(z_i), & \text{if } x_i \text{ is textual} \end{cases}, \tag{28}$$

where x_i denotes a token in the input sequence (textual or visual). Q, K, and V are the concatenated query, key, and value matrices from all Q_i, K_i, and V_i, respectively. z_i represents the intermediate state after residual connection and layer normalization on the shared self-attention output.

LMFusion initializes its language module from LLaMA-3 [236], freezing language parameters and fine-tuning only the vision module. This preserves language proficiency while enabling visual understanding and generation, obviating the need for large-scale retraining. To further improve flexibility and efficiency, X-Fusion [232] introduces a vision expert into a frozen pretrained LLM via a modular design, supporting new modalities without compromising language ability and significantly reducing computational cost. This approach is highly scalable and readily extends to modalities such as audio. BAGEL [22] adopts a Mixture-of-Transformers (MoT) architecture with two specialized experts: one for understanding (using NTP) and one for generation (using Rectified Flow). These experts share self-attention layers to facilitate seamless information exchange and synergy between understanding and generation. Mogao also employs an MoE structure, where a shared self-attention layer processes interleaved text and visual tokens, and modality-specific FFNs and attention projectors act as lightweight experts. Text and ViT-derived visual features are routed to the text branch, while VAE-generated tokens are directed to a separate generation branch. Timestep embeddings modulate visual features via AdaLN [16] to support rectified flow-based generation. Both BAGEL and Mogao [24] further adopt dual visual encoders to capture multi-level visual features, mitigating interference between understanding and generation tasks.

(3) Pros and Cons

Autoregressive-Diffusion Hybrid Modeling exhibits distinct advantages and limitations in UFM development. This paradigm enhances visual generation instruction-following capabilities by integrating autoregressive and diffusion mechanisms, while avoiding the information transmission bottlenecks characteristic of Modular Joint Modeling, thus achieving superior generation consistency and architectural coherence. Moreover, the incorporation of diffusion modeling yields substantially higher visual generation quality compared to purely autoregressive approaches. However, the framework faces several challenges. The requisite noise injection during denoising-based learning may compromise understanding performance, while parameter sharing between autoregressive and diffusion objectives introduces potential training conflicts, significantly increasing computational complexity and overhead. Despite these limitations, this research direction presents a promising pathway for unified discrete-continuous modality modeling, offering viable solutions for developing highly generalizable multimodal understanding and generation systems.

3.3.4 *Other Types*

(1) Definition

In addition to the widely adopted decoder-only LLM backbone for End-to-End Unified Modeling, a number of works [123], [97], [138], [163] have explored alternative backbone architectures. Representative examples include encoder–decoder transformers and State Space Models (e.g., Mamba [246]). These approaches introduce different modeling perspectives, bringing distinctive characteristics in terms of model design, training dynamics, and capability trade-offs. As such, they serve as valuable complements to

the dominant paradigms and are consolidated here under the "Other Types" category, providing exploratory directions for advancing End-to-End Unified Modeling.

(2) Relevant Works

Encoder-Decoder Architecture. Benefiting from the powerful modeling capabilities of the transformer architecture [247], early efforts in unified modeling typically adopted the encoder-decoder framework, with a common strategy of reformulating diverse tasks into a unified sequence-to-sequence (Seq2Seq) format.

Huang et al. [233] propose a single-transformer-based architecture that supports both text-to-image and image-to-text tasks within a unified model, eliminating the resource overhead and architectural complexity of using two separate models. Their approach introduces a two-level feature encoding mechanism, where dense features are used for image captioning and discrete features for image generation, striking a balance between generation efficiency and semantic expressiveness. OFA [123] introduces the principles of "Task-Agnosticism", "Modality-Agnosticism", and "Task-Comprehensiveness". It unifies text, image, and location tokens into a shared vocabulary and adopts an instruction-based learning paradigm for both pretraining and fine-tuning stages. This enables the model to handle a variety of unimodal and cross-modal tasks, such as image captioning and visual grounding, without relying on task-specific modules and demonstrates strong generalization capabilities. Davinci [234] extends the prefix language modeling approach to multimodal scenarios by proposing the prefix multi-modal modeling framework. It partitions image-text pairs into prefixes and suffixes, and leverages a cross-modal prefix-suffix modeling mechanism. This allows the model to learn both image and text generation in a self-supervised manner, efficiently utilizing large-scale image-caption paired data, and achieving a simple yet powerful multimodal modeling capability.

The Unified-IO series [97], [138] extends this unified modeling paradigm by converting all inputs and outputs into discrete token sequences for Seq2Seq learning, while significantly expanding task coverage to include depth estimation, keypoint detection, and other vision tasks. The follow-up Unified-IO-2 [138] further extends modality coverage to include text, image, audio, and action, and addresses the challenges of multimodal unified training through architectural enhancements such as 2D rotary positional embeddings, QK normalization, and scaled cosine attention, thereby improving both the modeling capacity and the training stability.

State Space Models. In the pursuit of lightweight and structurally innovative unified multimodal modeling, OmniMamba [163], built upon the Mamba-2 [248] architecture, successfully circumvents the quadratic computational complexity brought by traditional transformer architectures, while simultaneously improving modeling efficiency. OmniMamba adopts a decoupled visual representation strategy, incorporating SigLIP [33] and DINOv2 [249] to extract visual features for visual understanding, and leverages the tokenizer trained with LlamaGen [239] to discretize images into tokens for autoregressive image generation. To improve adaptability across different tasks, OmniMamba integrates task-specific LoRA modules into the input linear projection of each Mamba-2 layer. Coupled with a proposed two-stage decoupled training strategy, the model achieves competitive performance using only around 2 million image-text pairs.

Graph. Targeting the strong semantic correlations between images and text, GraphGPT-o [235] proposes representing multimodal inputs as a Multimodal Attributed Graph (MMAG) and leverages graph structures for content modeling, enabling the capture of complex relationships between cross-modal entities. To address the graph scalability explosion introduced by this structure, GraphGPT-o incorporates PageRank, effectively controlling graph complexity while preserving key information.

4 Encoding

In the encoding stage of UFMs, to achieve effective multimodal understanding and generation input modalities such as images, videos, and audio need to be transformed into suitable latent representations, enabling seamless alignment and integration with the text-based representations inherent in LLMs. Depending on how modality information is encoded and represented, we categorize encoding methods into three distinct types: Continuous (Sec. 4.1), Discrete (Sec. 4.2), and Hybrid (Sec. 4.3). In the following sections, we take the image modality as the primary modality for explanation, followed by complementary descriptions of the video and audio modalities. Tab. 2 outlines the encoding modules employed for different modalities in recent works, and Fig. 7 overviews the encoding strategies of UFMs, grouped into Continuous, Discrete, and Hybrid.

4.1 Continuous Representation

Continuous representation involves encoding multimodal inputs as differentiable, real-valued vector sequences, thereby enabling seamless semantic alignment across modalities within UFM frameworks. In this section, we systematically review representative paradigms for continuous encoding across different modalities, including image (Sec. 4.1.1), video (Sec. 4.1.2), and audio (Sec. 4.1.3).

4.1.1 Image

Continuous image encoding in UFMs builds upon established encoders from MLLMs, transforming images into differentiable, real-valued vector sequences that facilitate seamless semantic alignment with language models while minimizing information loss. Such representations are particularly advantageous for semantically demanding tasks, including visual question answering, cross-modal retrieval, and multimodal reasoning. Below, we summarize representative continuous image encoding approaches, such as VAE-based models, CLIP ViT, and Q-Former architectures.

VAE-Based Paradigm. The first continuous encoding approach leverages Variational Autoencoders (VAEs) [93] and σ-VAE [250]. VAEs compress images into continuous latent vectors using an encoder, subsequently reconstructing the original image pixels through a decoder. This methodology provides a high compression ratio, effectively avoiding

TABLE 2: Typical Encoding modules for the generation of UFM based on representation type and generative modality. In this context, **General** refers to the general tokenizer or feature extractor used as the source model for the encoding module of UFM. **Advanced** and **Unified** denote UFM's encoding modules featuring notable design, detailed in Sec. 4.

Model	Type	Year	Modality	Architecture	Source Model	VQ Type	Max Resolution	Notable Design
General								
VAE [93]	continuous	2013	image	VAE	-	-	256 × 256	-
σ-VAE [250]	continuous	2020	image	VAE	-	-	256 × 256	-
CLIP [31]	continuous	2021	image	ViT	-	-	336 × 336	-
EVA_CLIP [53]	continuous	2023	image	ViT	-	-	448 × 448	-
BLIP-2 [196]	continuous	2023	image	ViT	-	-	384 × 384	-
NaViT [251]	continuous	2023	image	ViT	-	-	896 × 896	-
SigLIP [33]	continuous	2023	image	ViT	-	-	384 × 384	-
UniT [252]	continuous	2024	image	ViT	-	-	896 × 896	-
VideoSwin [253]	continuous	2021	video	SwinTransformer	-	-	224 × 224	-
ViViT [254]	continuous	2021	video	ViT	-	-	224 × 224	-
SAN-M [255]	continuous	2020	audio	Transformer	-	-	-	-
AST [256]	continuous	2021	audio	Transformer	-	-	-	-
CLAP [257]	continuous	2022	audio	Transformer	-	-	-	-
MERT [258]	continuous	2023	audio	CNN + Transformer	-	-	-	-
ImageBind [39]	continuous	2023	audio, image, video	ViT	-	-	224 × 224	-
VQVAE [94]	discrete	2017	image	VAE	-	VQ	256 × 256	-
VQGAN [96]	discrete	2020	image	VAE	-	VQ	256 × 256	-
RQVAE [259]	discrete	2022	image	VAE	-	RQ	256 × 256	-
MoVQGAN [260]	discrete	2022	image	VAE	-	MoVQ	1024 × 1024	-
SpecVQGAN [261]	discrete	2021	audio	VAE	-	VQ	-	-
Encodec [262]	discrete	2022	audio	VAE	-	VQ	-	-
SpeechTokenizer [263]	discrete	2023	audio	VAE	-	VQ	-	-
Advanced								
MiniGPT-5 [134]	continuous	2023	image	ViT, U-Net	BLIP2, SD	-	224 × 224	Generative Vokens
MM-Interleaved [139]	continuous	2024	image	ViT	CLIP	-	224 × 224	Multi-Modal Feature Synchronizer
Janus-Flow [152]	continuous	2024	image	ViT, U-Net(VAE)	SigLIP, SDXL_VAE	-	384 × 384	Dual Branch
Mogao [24]	continuous	2025	image	CNN, ViT	SigLIP, VAE	-	512 × 512	Dual Branch
BAGEL [22]	continuous	2025	image	CNN, ViT	SigLIP2, FLUX_VAE	-	980 × 980	Dual Branch
UniFluid [211]	continuous	2025	image	CNN, ViT	SigLIP,VAE	-	256 × 256	Dual Branch
UniWorld-V1 [200]	continuous	2025	image	ViT	SigLIP, Qwen2.5-VL	-	512 × 512	Dual Branch
Pisces [201]	continuous	2025	image	ViT	EVA-CLIP, SigLIP	-	336 × 336	Dual Branch
show-o2 [175]	continuous	2025	image, video	CNN + ViT	3D causal VAE	-	432 × 432	Dual Branch
SynerGen-VL [264]	discrete	2024	image	CNN	MoVQGAN	MoVQ	256 × 256	Token Folding
EMU3 [25]	discrete	2024	video, image	3D-CNN	MoVQGAN	MoVQ	512 × 512	3D Tokenizer
show-o [116]	mix	2024	image	CNN	MAGVIT-v2	LFQ	256 × 256	Casual 3D CNN
OmniMamba [163]	mix	2025	image	ViT, CNN	SigLIP, DINOv2, LlamaGen	VQ	384 × 384	Dual Branch
UniToken [165]	mix	2025	image	ViT, CNN	SigLIP, VQGAN	VQ	768 × 768	Dual Branch
Video-LaViT [141]	mix	2024	video	CNN + ViT	LaViT	VQ	224 × 224	Motion-Specific Tokenizer
FOCUS [265]	mix	2025	image, video	ViT, CNN	CLIP,MoVQGAN	SimVQ	768 × 768	Dual Branch
Unified								
MetaMorph [155]	continuous	2024	image	ViT	SigLIP	-	512 × 512	Image Alignment
PUMA [151]	continuous	2024	image	ViT	CLIP	-	256 × 256	Multi-Granular Representations
BLIP3-o [23]	continuous	2025	image	ViT	Qwen2.5-VL	-	3584 × 3584	Image Alignment
Nexus-Gen [168]	continuous	2025	image	ViT	Qwen2.5-VL	-	3584 × 3584	Image Alignment
QLIP [160]	continuous	2025	image	ViT	EVA_CLIP	-	392 × 392	Contrastive Loss
Emu2 [137]	continuous	2023	image, video	ViT	EVA_CLIP	-	448 × 448	Image Alignment
SEED [130]	mix	2024	image	ViT	BLIP-2	VQ	384 × 384	Causal Q-Former + CL
LaViT [133]	mix	2024	image	ViT	CLIP	VQ	336 × 336	Dynamic Visual Tokenization
ILLUME [164]	mix	2024	image	ViT	UNIT	VQ	448 × 448	Image Alignment
VILA-U [149]	mix	2024	image	CNN, ViT	CLIP, RQVAE	ResVQ	384 × 384	CL + ReCon Loss
TokenFlow [154]	mix	2024	image	ViT, CNN	SigLIP, VQGAN	MsVQ	384 × 384	Dual Branch
MUSE-VL [215]	mix	2024	image	ViT, CNN	SigLIP, VQGAN	VQ	384 × 384	Dual Branch
ILLUME+ [203]	mix	2025	image	ViT, CNN	CLIP,MoVQGAN	SimVQ	512 × 512	Dual Branch
DDT [204]	mix	2025	image	ViT, Transformer	SD3 or VIT	VQ	256 × 256	Discrete Diffusion Timestep Tokenization
Tar [266]	mix	2025	image	ViT	SigLIP	VQ	384 × 384	Dual Branch
SemHiTok [216]	mix	2025	image	ViT	SigLIP	SGHC VQ	384 × 384	Dual-Codebook
TokLIP [170]	mix	2025	image	ViT, CNN	CLIP, VQGAN	VQ	384 × 384	Dual Branch
UniTok [159]	mix	2025	image	CNN	ViTamin	MCQ	256 × 256	Multi-Codebook Quantization
LM-MSN [207]	mix	2025	audio	CNN	Stable audio open	VQ	-	Audio Alignment

discrete quantization-related information loss. σ-VAEs introduce a variance-preserving regularizer to prevent latent-space collapse, ensuring latents remain suitable for autoregressive or diffusion-based generation.

For instance, TransFusion [117] simultaneously trains text next-token prediction and image VAE-latent diffusion noise prediction within a unified Transformer architecture, enabling seamless text-image joint generation. Similarly, OmniFlow [227] expands VAE latent vectors to text, image, and audio streams, using cross-modal joint attention in its Omni-Transformer to support arbitrary modality translation. LatentLM [221] employs σ-VAE continuous latent vectors integrated with next-token diffusion alongside textual tokens, significantly reducing inference steps.

CLIP ViT-Based Paradigm. The second continuous encoding approach directly utilizes CLIP [31] and their vari-

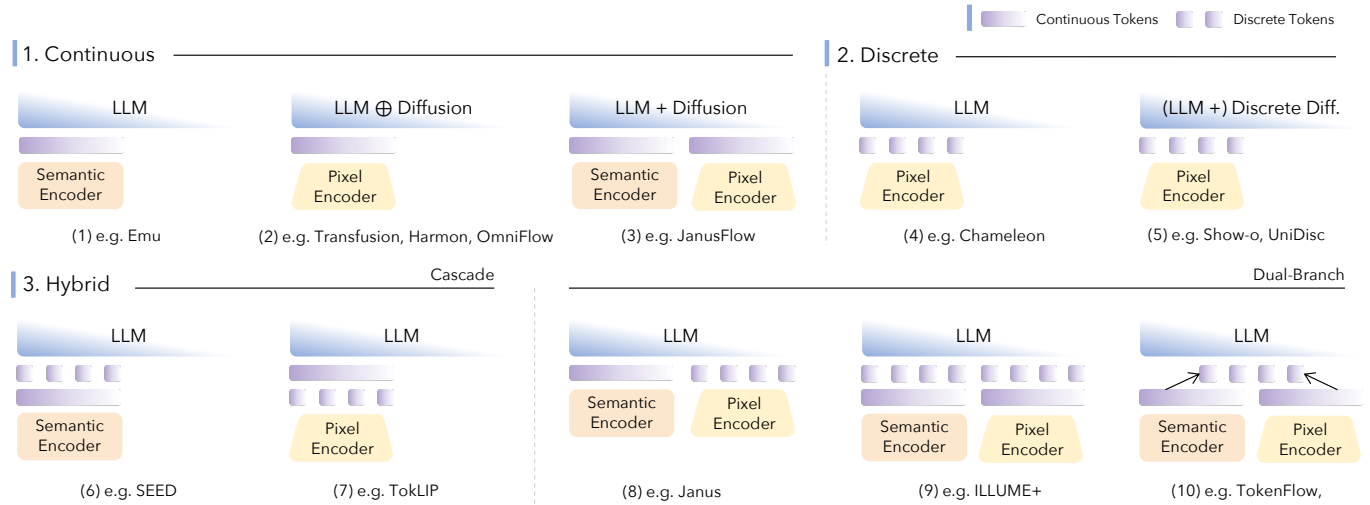

Fig. 7: Typical Encoding Strategies of UFM. Encoding Strategies of UFM are divided into 3 categories: ***Continuous*** (Sec. 4.1), ***Discrete*** (Sec. 4.2), and ***Hybrid***(Sec. 4.3), based on latent representation type. The encoding module, latent representation, and backbone are illustrated in each category, with annotations referring to classic UFM methods.

ants as image encoders. In this paradigm, Images pass through a pre-trained or trainable CLIP ViT to produce high-dimensional continuous patch embeddings, which are then mapped into the LLM embedding space via a learnable linear layer. Several improved versions of CLIP have been proposed to enhance encoding efficiency and performance, such as EVA-CLIP [53], which inherits stronger vision features from large-scale EVA pre-training, and SigLIP [33], which replaces the softmax contrastive loss with a sigmoid variant, improving stability and throughput on web-scale data, and a series of UFMs leverages these models as their visual encoder backbone to exratct semantic information. For instance, Emu2 [137] employs EVA-02-CLIP-E-plus, initializing its visual encoder with this powerful pre-trained model to extract continuous embeddings from images. UniFluid [211] integrates the SigLIP image encoder to produce continuous visual tokens that seamlessly combine with textual tokens in a unified autoregressive sequence. MetaMorph [155] encodes images with the SigLIP vision encoder, yielding continuous visual embeddings, which are interpolated to a fixed token size and projected into the LLM's dimension through a trainable layer. This enables the model to autoregressively predict visual tokens as outputs, effectively bridging visual understanding and generation within a unified multimodal framework.

CLIP + Q-Former Paradigm. The third continuous encoding approach integrates the Q-Former architecture, initially introduced in BLIP-2 [196], exemplified by UFMs like SEED-LLaMA [135], to achieve more efficient extraction and semantic alignment of visual information with LLMs. This paradigm begins by extracting initial multi-scale or multi-level visual features from images using a frozen pre-trained visual encoder. It then introduces a trainable Q-Former module containing a set of learnable query embeddings. These embeddings interact via cross-attention with the frozen visual encoder's rich feature outputs, actively querying and extracting a fixed number of compact visual representations most relevant to the latent textual information or specific task requirements.

Pros and Cons. The VAE-based paradigm excels at preserving low-frequency information and global structure, making it advantageous for tasks requiring pixel-level fidelity or holistic reconstruction. CLIP ViT-based approaches, benefiting from large-scale image-text contrastive pretraining, produce features that are well-aligned with textual semantics and are effective for high-level understanding and reasoning. Q-Former-based methods further distill and compress visual representations, acting as efficient adapters that reduce redundancy and enhance semantic alignment with language models.

However, each approach has inherent limitations. VAE-based methods, while strong in detail preservation, often lack semantic abstraction and underperform on tasks demanding high-level understanding. CLIP ViT-based features, though semantically rich, are less effective at capturing fine-grained details and spatial relationships, and cannot directly support image reconstruction. Q-Former performance is contingent on the quality of the underlying visual encoder; suboptimal initial features limit its effectiveness. These trade-offs highlight the ongoing challenge of balancing semantic richness, detail fidelity, and generative flexibility in continuous image encoding.

4.1.2 *Video*

Continuous video encoding maps videos—rich in spatiotemporal dynamics—into continuous vector representations suitable for LLMs. Unlike static images, it should capture temporal variation and spatiotemporal interactions. Early approaches independently processed each frame using static encoders like CLIP ViT, followed by temporal aggregation methods such as averaging, concatenation, or RNN-based aggregation. However, such methods inadequately model explicit motion trajectories and cross-frame interactions, limiting performance on tasks involving action recognition, event reasoning, and long-range dependencies.

To address these limitations, recent work adopts Video Swin Transformer [253] and ViViT [254]. Video Swin Transformer segments videos into non-overlapping 3D patches,

encoding these tokens using hierarchical transformers with 3D shifted-window self-attention. This design efficiently captures local spatiotemporal dependencies and aligns encoded video features with textual embeddings via contrastive learning methods, such as VALOR [267]. However, information capturing long-range temporal dependencies propagates incrementally through multiple hierarchical layers via shifted windows, potentially limiting long-range context modeling efficiency. In contrast, ViViT transforms videos into spatiotemporal token sequences through uniform frame sampling, independently tokenizing each frame into patches, followed by linear projection and concatenation across frames. Transformer encoders continuously encode these tokens, and the resulting video features are mapped to the LLM's embedding space via dedicated Video Understanding Adapters, such as MuMu-LLaMA [157]. Despite its comprehensive modeling, ViViT's cost scales sharply with frame count and resolution, challenging longer or high-resolution videos.

4.1.3 Audio

The continuous encoding of audio modality aims to transform complex acoustic signals into sequences of continuous vectors that can be effectively understood and leveraged by LLMs. Unlike visual data, audio inherently involves intricate temporal dynamics and frequency-domain characteristics, posing unique challenges in extracting meaningful features for multimodal alignment.

One representative continuous audio encoding approach is the Audio Spectrogram Transformer (AST) [256], as adopted in VALOR [267]. AST converts acoustic signals into Mel spectrograms, segments them into patches, and feeds them to a standard Transformer encoder. Leveraging self-attention mechanisms, AST implicitly captures intricate time-frequency relationships from raw spectrogram patches without explicit temporal or spectral modeling constraints. In VALOR, the resulting general-purpose audio embeddings are contrastively aligned with visual and textual embeddings in a shared semantic space.

M2-Omni [192] introduces the SAN-M [255] architecture optimized specifically for audio encoding. Recognizing the limitations of pure self-attention mechanisms in capturing fine-grained local temporal contexts, SAN-M integrates Deep Feedforward Sequential Memory Networks (DFSMN) into the standard Transformer framework. DFSMN explicitly enhances the encoding of short-range temporal dependencies through efficient static memory modules, complementing the dynamic long-range dependency capturing ability of self-attention. In M2-Omni [192], the SAN-M-generated continuous audio embeddings are subsequently mapped via an MLP into the multimodal shared embedding space for cross-modality tasks. MuMu-LLaMA [157] utilizes a large-scale self-supervised pretrained model MERT [258] for comprehensive music understanding, and the pretrained MERT features are further aligned to the LLM input space through a learnable adapter, enabling effective semantic integration of audio modality representations within the unified multimodal framework.

4.2 Discrete Representation

Discrete representation refers to the process of mapping multimodal inputs into sequences of discrete tokens, typically through quantization techniques such as vector quantization, and facilitates unified multimodal modeling and generation. In this section, we outline the core concepts of discrete encoding and introduce methods for image, video, and audio modalities. For images, we discuss common paradigms based on vector quantization, including VQ-VAE, VQ-GAN, and multi-level quantization strategies (Sec. 4.2.1). We then cover discrete encoding approaches for video (Sec. 4.2.2) and audio (Sec. 4.2.3) modalities.

4.2.1 Image

Discrete encoding typically leverages Vector Quantization (VQ), initially mapping images into a continuous latent space and subsequently quantizing these vectors by finding their nearest neighbors within a learnable codebook, replacing them with corresponding discrete indices. The discrete encoding of image modalities aims to convert continuous visual information into discrete visual tokens, a representation naturally compatible with the discrete input format of LLMs. For instance, UniCode [268] integrates VQ-generated discrete indices directly into the LLM vocabulary, enabling visual and textual modalities to share a unified codebook.

Common VQ approaches include VQ-VAE [94], as exemplified by Liquid, which employs VQ-VAE as a visual tokenizer, directly embedding discrete image tokens into the shared vocabulary of a decoder-only LLM, simultaneously supporting both visual understanding and generation tasks. VQ-GAN [96] variants, such as those employed by Unified-IO [97], Divter [184], RA-CM3 [213], and LWM [142], enhance VQ-VAE by incorporating adversarial discriminator networks and GAN losses, significantly improving perceptual quality and realism in generated images. Both VQ-VAE and VQ-GAN essentially follow the same core single-step quantization mechanism.

Unified-IO adopts VQ-GAN as a universal visual tokenizer, discretizing dense visual structures (images, segmentation masks, depth maps) into token sequences for LLM input. In contrast, multi-level Residual Vector Quantization (RQ-VAE [259]), utilized in models like VILA-U [149], employs an iterative quantization strategy. RQ-VAE features multiple cascaded quantizers and codebooks, each capturing residual information left unencoded by preceding quantizers, allowing the accumulation or combination of outputs to represent original latent features with finer granularity and higher efficiency. This approach enhances the model's ability to capture detailed visual information.

Another advancement is MoVQ, introduced by MoVQ-GAN [260], which employs multi-channel quantization. Without increasing the size of individual codebooks, MoVQ simultaneously maps each image patch into multiple codebook indices, enhancing reconstruction quality and generation speed via parallel sampling techniques. EMU3 [25] leverages MoVQ to discretize both images and video segments into multi-channel tokens, enabling a single transformer model to perform high-fidelity image and video generation, as well as visual-language understanding, purely through next-token prediction. Additionally,

TokenFlow [154] leverages Multi-Scale Vector Quantization (MSVQ) from VAR [115] to encode images into discrete feature maps across multiple scales and generate tokens autoregressively in a coarse-to-fine manner—lower-resolution codebooks capture global structure, while higher-resolution codebooks preserve local details.

Pros and Cons. By converting continuous latent features into compact discrete tokens, these VQ methods enable direct compatibility with the token-based input of language models. Approaches such as VQ-VAE and VQ-GAN provide efficient compression and allow both visual and textual information to share a unified codebook, supporting joint visual understanding and generation. RQ-VAE and MoVQ enhance representation flexibility, capturing both global structure and fine local details by leveraging multi-stage or multi-channel quantization. Additionally, MSVQ strategies enable models to encode hierarchical image information, improving both perceptual quality and generative accuracy.

Despite their advantages, discrete image encoding methods face several challenges. A fundamental trade-off exists between optimizing tokens for faithful image reconstruction and ensuring rich semantic abstraction. Tokens tailored for pixel-level fidelity often lack the semantic depth required for high-level reasoning and robust alignment with language models. The effectiveness of these methods is further constrained by codebook design and quantization strategies, which may limit adaptability to diverse data distributions. Moreover, while multi-stage or multi-channel quantization improves detail retention, it increases computational complexity and risks redundancy. Consequently, discrete encoding approaches need to carefully balance compression efficiency, semantic alignment, and computational overhead in unified multimodal frameworks.

4.2.2 Video

Discrete video encoding transforms continuous spatiotemporal video signals into discrete token sequences, enabling unified autoregressive modeling with textual inputs. This approach converts video data into a format compatible with multimodal backbones, facilitating seamless multimodal integration and joint processing.

For instance, Emu3 [25] leverages MoVQGAN [260] to discretely encode videos on a frame-by-frame basis, quantizing the spatiotemporal features into discrete visual tokens. These visual tokens are directly interleaved with textual tokens and jointly fed into a unified autoregressive Transformer model. This framework enables pure next-token video understanding and generation without relying on diffusion-based models, significantly enhancing the efficiency and coherence of multimodal modeling. In comparison, MIO [202] introduces a dynamic frame-sampling strategy, employing SEED-Tokenizer [130] to discretize each selected frame independently into visual tokens. Subsequently, these tokens are sequentially concatenated with text tokens and processed autoregressively by an LLM, effectively supporting a unified modeling scheme for multimodal tasks. Similarly, LWM [142] discretizes videos using VQ-GAN [96], encoding each frame individually into discrete visual tokens. These tokens are concatenated temporally along with textual tokens, serving as inputs to the LLM for joint multimodal modeling. This method efficiently captures both visual and temporal information in a discrete token format that aligns seamlessly with the generative capabilities of autoregressive language models.

4.2.3 Audio

Discrete encoding methods for the audio modality aim to represent complex acoustic signals as discrete tokens, thus enabling unified multimodal understanding and generation within language models. Recent research typically employs hierarchical quantization methods to effectively capture the multi-scale acoustic features inherent in audio signals.

For example, C3LLM [206] utilizes EnCodec [262] to perform hierarchical discrete quantization on audio signals, encoding audio features at multiple layers into discrete codebook indices. These indices are then incorporated directly as "audio tokens" into the vocabulary of the LLM, achieving unified modeling of audio signals alongside textual inputs. Similarly, AnyGPT [140] adopts the EnCodec framework for hierarchical residual quantization of audio inputs, including speech and music. Audio segments are represented as discrete tokens derived from multiple codebooks, and these tokens serve as specialized audio vocabulary entries within the LLM. Such an approach facilitates comprehensive multimodal integration and generation across audio and other modalities. Additionally, MIO [202] employs SpeechTokenizer [263] to hierarchically discretize speech signals, transforming them into multiple layers of discrete tokens. These speech tokens are then integrated as "speech vocabulary" within the LLM, facilitating seamless multimodal representation and enabling tasks that require joint processing and generation of speech, text, and visual data within a unified modeling paradigm.

4.3 Hybrid Representation

Hybrid representation strategies combine continuous and discrete encoding approaches to address their individual limitations. Continuous encodings excel at capturing semantic information for understanding tasks but lack the pixel-level detail necessary for high-quality generation. Discrete encodings preserve fine-grained details for reconstruction but often produce tokens poorly aligned with textual embeddings, limiting understanding performance. To overcome these challenges, recent UFMs increasingly adopt hybrid encoding to achieve comprehensive visual representations that simultaneously support complex reasoning and high-fidelity generation. These approaches integrate semantic richness from continuous features with detailed reconstruction capabilities from discrete tokens, bridging the distinct requirements of understanding and generation tasks within unified frameworks. Current hybrid encoding methodologies follow two primary architectures: cascade structures (Sec. 4.3.1) that sequentially process features through alternating continuous and discrete representations, and dual-branch structures (Sec. 4.3.2) that employ parallel encoding paths—one optimized for semantic understanding and another for detailed generation—subsequently integrating these features at the tokenizer or backbone level.

4.3.1 Cascade Encoding Strategy

Cascade hybrid encoding approaches seek to integrate visual understanding and generation capabilities within a

unified pathway by combining continuous and discrete representations sequentially. The key idea is to maintain a unified visual representation that simultaneously contain semantic information (for understanding) and detailed pixel-level information (for reconstruction).

An early example is SEED [130], which first reshapes continuous 2D visual features into 1D causal semantic embeddings via a Causal Q-Former, then discretizes these embeddings using a VQ codebook. This "semantic-to-discrete" cascade strategy effectively aligns with the autoregressive nature of LLMs but introduces additional transformation modules, complicating the tokenizer architecture. To improve training efficiency, ILLUME [164] simplifies this strategy by directly quantizing continuous visual features in a predefined semantic space, thereby achieving effective semantic alignment with fewer architectural complexities.

VILA-U [149] employs a different cascade approach. It first encodes images into continuous semantic features using a pretrained ViT, subsequently converting these features into discrete tokens via residual quantization. These discrete tokens are then directly used for contrastive learning with text embeddings. Despite its efficiency, VILA-U [149] faces challenges in simultaneously optimizing reconstruction and contrastive losses, leading to potential conflicts and convergence issues. Addressing these limitations, UniTok [159] argues that the core issue lies in the limited representational capacity of traditional discrete spaces. To overcome this, UniTok first projects continuous features via attention modules and then applies Multi-Codebook Quantization (MCQ). MCQ exponentially enlarges the effective discrete vocabulary size, fundamentally enhancing the representational power of discrete spaces, thus making joint optimization of contrastive and reconstruction losses practical and efficient.

The Tar [266] introduces a "text-aligned visual tokenizer" strategy, extracting high-level visual features using SigLIP2 [269], and then obtaining multi-scale visual details via Scale-Adaptive Pooling. These features are discretized using a text-aligned codebook initialized from LLM embeddings, naturally aligning visual tokens containing both semantic information and visual details within the LLM's textual embedding space. Similarly, UniCode^2 [268] clusters visual features obtained from SigLIP to create a frozen semantic codebook, encoding images by table lookup and adding small trainable offsets to each token, producing semantic-aligned discrete visual tokens. In contrast to the prevalent semantic-first strategy, TokLIP [170] adopts a "discrete-to-continuous" encoding strategy, which discretizes images using VQGAN, followed by projecting discrete tokens into CLIP feature space via a trainable MLP, and subsequently extracting high-level continuous features through a ViT.

4.3.2 Dual-Branch Hybrid Encoding Strategy

Dual-branch hybrid encoding strategies explicitly decouple visual feature extraction for understanding and generation tasks through parallel architectural pathways. This approach employs two independent encoding branches: one optimized for semantic feature extraction to support understanding tasks, and another designed for detailed pixel-level encoding to enable generation capabilities. These parallel feature streams are subsequently integrated through fusion mechanisms at various stages—within the tokenizer, prior to LLM input, or within the LLM backbone itself—to achieve unified multimodal modeling. Current dual-branch methodologies can be categorized into two primary architectures: (1) approaches that fuse parallel feature streams into unified representations before the model core processing, and (2) approaches that maintain completely separate encoding pathways for understanding and generation tasks without early-stage feature integration.

Fusion into a Unified Representation. In the first type of dual-branch methods, the model simultaneously extracts semantic and pixel-level information through parallel encoding branches, then integrates these distinct feature representations via concatenation or fusion to form a single, unified visual input for the backbone.

For example, TokenFlow [154] introduces a dual-codebook mechanism that utilizes shared discrete indices. Rather than directly fusing features, it aims to associate both semantic and pixel-level information within a single discrete index. During the VQ process, TokenFlow seeks a unique shared index for each image patch that minimizes a weighted distance sum between the patch's features and corresponding entries in the semantic and pixel codebooks. This indexing strategy enables individual discrete tokens to simultaneously encode high-level semantic concepts and fine-grained pixel-level details. MUSE-VL [215] presents the SDE-Tokenizer, which explicitly fuses continuous visual features from both an image pixel encoder and an external semantic encoder prior to vector quantization. The integrated features are subsequently discretized through a shared quantization process, effectively injecting semantic information into discrete tokens. This approach notably alleviates the potential optimization conflicts between contrastive semantic learning and reconstruction losses encountered in earlier models such as VILA-U.

UniToken [165] proposes an "information maximization" fusion strategy at the input stage, directly concatenating continuous semantic features extracted by SigLIP ViT with discrete token embeddings obtained from the Chameleon VQ-Tokenizer. This explicit concatenation preserves both semantic and pixel-level information comprehensively, thereby enhancing the representational richness and effectiveness of the combined visual features. In another distinct approach, SemHiTok [216] introduces a Semantic-Guided Hierarchical Codebook (SGHC) encoding mechanism. Initially, SemHiTok trains a semantic branch to perform high-level semantic quantization. Subsequently, for each semantic code, it trains additional child codebooks specifically designed to capture detailed pixel-level information correlated with that semantic category. As a result, each image patch is represented jointly by one semantic token and one pixel token. Despite its strong representational capability, SemHiTok faces scalability challenges; the overall vocabulary size scales quadratically with hierarchical layers, causing significant storage overhead and computational complexity, thus constraining its practical applicability.

Separated Dual Feature Paths. The second type of dual-branch structure adapts a fully separates feature encoding approach, using independent visual encoders and pipelines for the visual understanding and generation task. For example, Janus [150] designs distinct visual encoders for each task, utilizing a SigLIP ViT to extract continuous semantic

features for high-level visual understanding, and employing a VQGAN to obtain latent or discrete tokens for image generation. This complete disentanglement is intended to minimize feature conflicts between tasks, thereby optimizing the performance of each individual branch.

Building on this framework, VARGPT-v1.1 [167] introduces a multi-scale discrete tokenizer (MCQ) that encodes images as discrete token sequences at multiple resolutions. During generation, the model first predicts low-resolution tokens to establish the global structure and composition, and then incrementally generates higher-resolution tokens to fill in details, achieving a coarse-to-fine, autoregressive control over the generation process.

Recent models diverge from the conventional approach of using CLIP/SigLIP for continuous semantic encoding and VQGAN for discrete pixel-level encoding, instead adopting continuous representations for both high-level semantics and low-level pixel details. JanusFlow [152] and BAGEL [22] both employ SigLIP ViT and SDXL/FLUX VAE as semantic and pixel encoders, respectively. However, JanusFlow maintains complete decoupling between these features for understanding and generation tasks, while BAGEL enables interaction between semantic and pixel tokens through a Mixture-of-Tokens (MoT) mechanism, thereby mitigating information loss inherent in fully decoupled architectures. Similarly, Mogao [24] utilizes ViT and VAE to extract semantic and pixel features separately, proposing a Deep-Fusion mechanism within the LLM to facilitate rich interaction between the two streams. Show-o2 [175] leverages a 3D Causal VAE for both image and video encoding, extracting global semantics via a Semantic Projector while preserving fine-grained details through an MLP. These dual feature paths are concatenated before input and subsequently mapped through RMSNorm and an MLP to produce a unified representation encompassing both semantic and pixel-level information.

While fully separated dual-path approaches provide task-specific optimization flexibility, they create a representational gap between understanding and generation tasks, particularly evident in architectures like Janus [150] and JanusFlow [152]. For tasks requiring simultaneous comprehension and generation (e.g., interleaved text-image generation), this separation causes information loss due to insufficient joint modeling of semantic and pixel-level features, thereby limiting the UFM's performance on cross-modal interactive tasks. To address this limitation, ILLUME+ [203] proposes DualViTok, a unified dual-branch tokenizer that simultaneously captures semantic and textural information. DualViTok employs pretrained semantic and pixel encoders to discretize each image patch into paired semantic and pixel tokens, which are concatenated to form comprehensive representations preserving both textual alignment and visual fidelity. This design enhances performance across understanding, generation, and editing tasks. Similarly, FOCUS [265] extracts multi-resolution semantics using Qwen ViT and ConvNeXt-L, aggregated via a Gated Cross-Attention Adapter to produce continuous representations integrating spatial and semantic information. For pixel-level encoding, FOCUS employs MoVoGAN-based quantization, concatenating semantic and continuous pixel features before LLM input for text alignment. The aligned pixel features are subsequently vector-quantized, ensuring coherent semantic integration within the model framework.

5 Decoding

In the decoding stage of UFMs, non-text modalities typically necessitate dedicated decoders to align with the LLM, facilitating the transformation from latent representations to the corresponding original modality outputs. Depending on the representation modeling, we divide the decoding strategies into three categories: ***Continuous*** (Sec. 5.1), ***Discrete*** (Sec. 5.2), and ***Hybrid*** (Sec. 5.3). In the following sections, we take the image modality as the primary modality for explanation, followed by complementary descriptions of the video and audio modalities. Tab. 3 provides an overview of the decoding module used across various modalities in recent researches.

5.1 Continuous Representation

Decoding strategies for continuous latent representation are typically based on diffusion models, which take the continuous-space latent features from LLM and iteratively denoise them to generate the corresponding modality data.

Constructing Continuous Latent Representations. The continuous latent representation output by the LLM can be obtained in different ways. *1. Instruction-Based Latent.* MM-Interleaved [139], OmniGen2 [176], Ovis-U1 [177] directly extract the hidden states corresponding to the multimodal instruction from the final layer of the LLM as the latent representation. *2. Autoregression Latent.* Emu [129], PUMA [151], *et al.* adopt an autoregressive approach, where the LLM is prompted to generate a fixed-length sequence of hidden states as the latent representation. Since latent representation tokens lack the recursive structure inherent in natural language, some methods avoid relying on preceding embeddings when decoding continuous features. *3.Special Token Prompting.* NExT-GPT [131], X-VILA [205], *et al.*, extend the text embedding layer by introducing fixed special tokens (*e.g.*, `<image>`) as inputs to LLM, and the hidden states corresponding to these tokens are used as latent representations. *4. Learnable Query Prompting.* DreamLLM [132], SEED-X [144], *et al.* introduce learnable queries as LLM inputs to enhance multimodal in-context learning capabilities [166], as shown in Fig. 8 (2). These approaches mitigate the accumulation of errors caused by inaccurate preceding embeddings during autoregressive decoding of continuous features [168], and further enable the use of bidirectional attention within the LLM to support parallel generation of latent representations.

We categorize continuous representation decoding methods into ***External Generation*** (Sec. 5.1.1) and ***Internal Generation*** (Sec. 5.1.2), according to the way in which the latent representation interfaces with the diffusion model.

5.1.1 External Generation

(1) Image

External generation methods typically align the latent representation via an additional connector and use it as the condition for the decoder, as illustrated in Fig. 8 (1)(2). In this setting, the decoder adopts complete diffusion models,

TABLE 3: Typical Decoding modules for the generation of UFM based on decoding strategy type and generative modality. In this context, ***General*** refers to the general generative model used as the source model for the decoding module of UFM. ***Advanced*** and ***Unified*** denote UFM's decoding modules featuring notable design, detailed in Sec. 5. ***Unified*** denotes decoding modules that are explicitly aligned with the corresponding encoders.

Model	Type	Year	Modality	Architecture	Source Model	Max Resolution	Notable Design
				General			
LDM [15]	continuous	2021	image	U-Net + VAE	-	512 × 512	-
SD 1.x [15]	continuous	2021	image	U-Net + VAE	-	512 × 512	-
SD 2.x [15]	continuous	2022	image	U-Net + VAE	-	768 × 768	-
InstructPix2Pix [270]	continuous	2022	image	U-Net + VAE	-	256 × 256	-
GLIGEN [271]	continuous	2023	image	U-Net + VAE	-	512 × 512	-
SDXL [27]	continuous	2024	image	U-Net + VAE	-	1024 × 1024	-
SD 3.x [103]	continuous	2024	image	DiT + VAE	-	1024 × 1024	-
Sana [272]	continuous	2024	image	DiT + VAE	-	4096 × 4096	-
FLUX.1 [14]	continuous	2024	image	DiT + VAE	-	1024 × 1024	-
Lumina-Image 2.0 [273]	continuous	2025	image	DiT + VAE	-	1024 × 1024	-
ZeroScope v2 [274]	continuous	2023	video	3D U-Net + VAE	-	1024 × 1576	-
I2VGen-XL [275]	continuous	2023	video	3D U-Net + VAE	-	720 × 1280	-
VideoCrafter 2 [276]	continuous	2024	video	3D U-Net + VAE	-	320 × 512	-
Wan [277]	continuous	2025	video	DiT + 3D VAE	-	720 × 720	-
AudioLDM [278]	continuous	2023	audio	U-Net + VAE	-	-	-
AudioLDM 2 [279]	continuous	2024	audio	U-Net + VAE	-	-	-
VQ-VAE [94]	discrete	2017	image, audio	VAE	-	128 × 128	-
VQ-GAN [96]	discrete	2020	image	VAE	-	256 × 256	-
RQ-VAE [259]	discrete	2022	image	VAE	-	256 × 256	-
Mo-VQGAN [260]	discrete	2022	image	VAE	-	256 × 256	-
Make-A-Scene [280]	discrete	2022	image	VAE	-	512 × 512	-
VAR [115]	discrete	2024	image	VAE	-	512 × 512	-
Inifnity [281]	discrete	2024	image	VAE	-	1024 × 1024	-
MAGVIT-v2 [282]	discrete	2024	image, video	3D VAE	-	512 × 512	-
EnCodec [262]	discrete	2022	audio	VAE + LSTM	-	-	-
SpeechTokenizer [263]	discrete	2023	audio	VAE	-	-	-
CosyVoice [193]	mix	2024	audio	DiT	-	-	-
				Advanced			
MiniGPT-5 [134]	continuous	2023	image	U-Net + VAE + TF	SD 1.x	512 × 512	TF Connector
GILL [194]	continuous	2023	image	U-Net + VAE + TF	SD 1.x	512 × 512	TF Connector
Ovis-U1 [177]	continuous	2023	image	DiT + VAE + TF	SD 3.x	1024 × 1024	Visual Prior
EasyGen [186]	continuous	2023	image	DiT + VAE + TF	UniDiffuser	512 × 512	TF Connector
MM-Interleaved [139]	continuous	2024	image	U-Net + VAE	SD 2.x	512 × 512	Visual Prior
Transfusion [117]	continuous	2024	image	VAE + U-Net	VAE	256 × 256	Token Upsampling
JanusFlow [152]	continuous	2024	image	VAE + ConvNeXt	SDXL	384 × 384	Token Upsampling
MetaQueries [166]	continuous	2025	image	DiT + VAE + TF	Sana	512 × 512	Visual Prior
OpenUni [173]	continuous	2025	image	DiT + VAE + TF	Sana	1024 × 1024	Visual Prior
Ming-Lite-Uni [169]	continuous	2025	image	DiT + VAE + TF	Sana	512 × 512	TF Conn., Multi-Tokens
UniWorld-V1 [200]	continuous	2025	image	DiT + VAE	FLUX.1	512 × 512	Visual Prior
OmniGen2 [176]	continuous	2025	image	DiT + VAE	Lumina-Image 2.0	1024 × 1024	Visual Prior
WeGen [199]	continuous	2025	image	U-Net + VAE	SDXL	1024 × 1024	Visual Prior
CoDi-2 [136]	continuous	2025	image	U-Net + VAE	SD 2.x	768 × 768	Visual Prior
MuMu-Llama [157]	continuous	2024	audio	U-Net + VAE	AudioLDM 2	-	TF Conn., Textual Loss
Vitron [191]	continuous	2024	image video	U-Net + VAE + TF 3D U-Net + TF	GLIGEN ZeroScope v2	512 × 512	TF Conn., Multi-Tokens, Textual Loss
NExT-GPT [131]	continuous	2023	image, video, audio	U-Net + VAE + TF U-Net + VAE + TF	SD 1.x , Zeroscope v2 AudioLDM	512 × 512	TF Conn., Textual Loss
X-VILA [205]	continuous	2024	image, video, audio	U-Net + VAE U-Net + VAE	SD 1.x, VideoCrafter 2 AudioLDM	512 × 512	Visual Prior
Spider [190]	continuous	2024	image, video, audio	U-Net + VAE + TF U-Net + VAE + TF	SD 1.x, ZeroScope v2 AudioLDM	512 × 512	MoE Conn.
Emu3 [25]	discrete	2024	image, video	3D VAE	MoVQGAN	720 × 720	3D Tokenizer
SynerGen-VL [264]	discrete	2024	image	3D VAE + TF	Emu3	512 × 512	Token Unfolding
Video-LaViT [141]	mix	2024	video	3D U-Net + VAE + TF	SVD	768 × 768	Motion Decomposition
FOCUS [265]	mix	2025	image, video	U-Net + VAE	SDXL	1024 × 1024	Visual Prior
				Unified			
PUMA [151]	continuous	2024	image	U-Net + VAE	SDXL	1024 × 1024	Image Align, Multi-Tokens
MetaMorph [155]	continuous	2024	image	U-Net + VAE	SD 1.x	512 × 512	Image Align
LatentLM [221]	continuous	2024	image	U-Net + VAE	VAE	384 × 384	Noise Perturbatiotion
SEED-X [144]	continuous	2024	image	U-Net + VAE	SDXL	1024 × 1024	Image Align, Visual Prior
Nexus-Gen [168]	continuous	2025	image	DiT + VAE	FLUX	1024 × 1024	Image Align
Pisces [201]	continuous	2025	image	U-Net + VAE	SDXL	1024 × 1024	Image Align
BLIP3-o [23]	continuous	2025	image	U-Net + VAE + DiT	Lumina-Next, SDXL	1024 × 1024	Image Align, Diff Conn.
Emu2 [137]	continuous	2023	image, video	U-Net + VAE	SDXL, SD 2.x	1024 × 1024	Image Align
VILA-U [149]	discrete	2024	image	VAE	RQVAE	384 × 384	Contrastive Loss
MUSE-VL [215]	discrete	2024	image	VAE	VQGAN	256 × 256	Semantic Loss
TokenFlow [154]	discrete	2024	image	VAE	VAR	384 × 384	Semantic Loss
SemHiTok [216]	discrete	2025	image	ViT	ViT-VQGAN	256 × 256	Semantic Loss
QLIP [160]	discrete	2025	image	ViT	BSQ-ViT	392 × 392	Contrastive Loss
UniTok [159]	discrete	2025	image	VAE	VQGAN	256 × 256	Contrastive Loss
TokLIP [170]	discrete	2025	image	VAE	Llama-Gen	384 × 384	Sem. Loss, Cont. Loss
SEED [130]	mix	2023	image	U-Net + VAE + TF	SD 1.x	512 × 512	TF Connector
LaViT [133]	mix	2024	image	U-Net + VAE + TF	LDM	1024 × 1024	Image Align, Sem. Dec.
ILLUME [164]	mix	2024	image	U-Net + VAE	SDXL	512 × 512	Image Align, Sem. Dec.
ILLUME+ [203]	mix	2025	image	U-Net + VAE	SDXL	1024 × 1024	Image Align, Sem. Loss
Tar [266]	mix	2025	image	DiT + VAE	Sana	1024 × 1024	Image Align
DDT [204]	mix	2025	image	DiT + VAE	SD 3.x	256 × 256	Image Align
LM-MSN [207]	mix	2025	audio	DiT + VAE	DiT	-	Audio Align

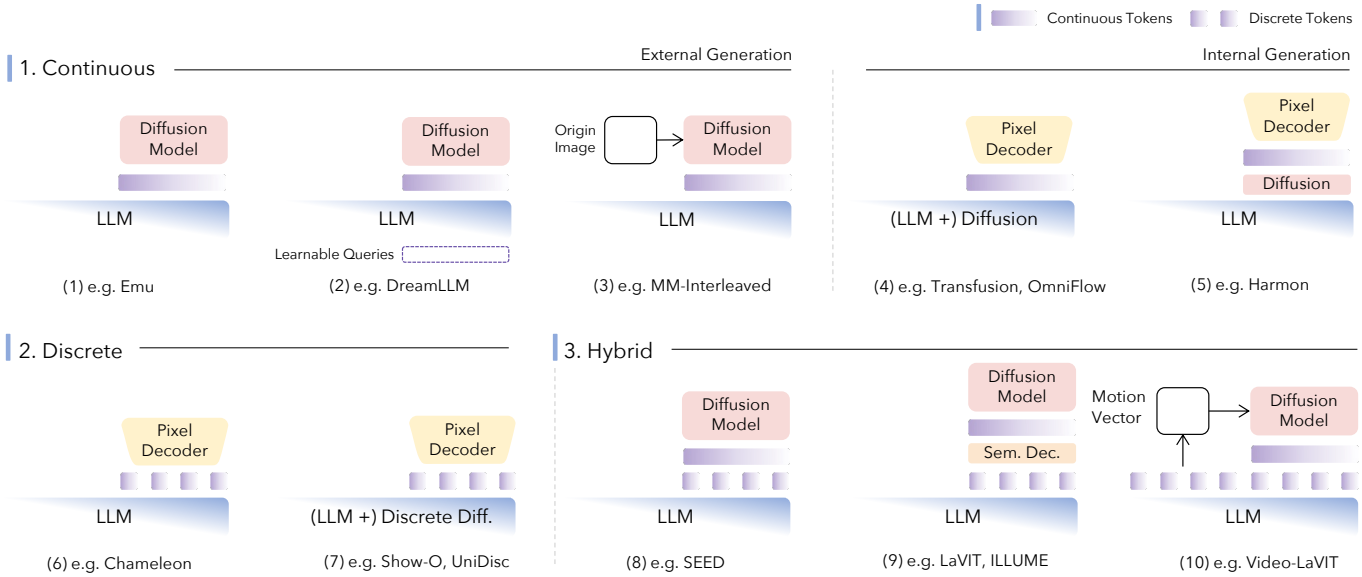

Fig. 8: Typical Decoding Strategies of UFMs. Decoding strategies of UFMs are divided into 3 categories: ***Continuous*** (Sec. 5.1), ***Discrete*** (Sec. 5.2), and ***Hybrid*** (Sec. 5.3), based on latent representation type. The decoding module, latent representation, and backbone are illustrated in each category, with annotations referring to classic UFM methods.

e.g., SD 1.x/2.x [15], SDXL [27], SD 3.x [103], Sana [272], FLUX.1 [14]. The generation of latent representations by the LLM is structurally decoupled from the denoising stage. It supports direct integration with widely used diffusion optimization methods, such as Classifier-Free Guidance (CFG) [283], as shown in Fig. 9 (4).

External generation methods can be further categorized according to the alignment strategy of the latent representation, namely ***Alignment with Text Space***, ***Alignment with Image Space***, and ***Dynamic Alignment***.

Alignment with Text Space. These alignment paradigms [136], [132] optimize the alignment by minimizing the diffusion loss between the generated image and the ground-truth image, while keeping the decoder frozen during training. They can be viewed as operating within the decoder's original text condition space, where the latent representation is learned through model-seeking strategies [132], and thus can be regarded as a generalized form of text-space alignment. MiniGPT-5 [134], NExT-GPT [131], EasyGen [186], *et al.*, introduce an additional Transformer-based connector to improve the alignment quality. Notably, NExT-GPT [131] and Vitron [191] incorporate textual loss, which minimizes the distance between the connector output feature and the decoder's textual representation to facilitate and accelerate the alignment process. GILL [194], VL-GPT [195], and Spider [190] completely remove the decoder from the training loop by training the Transformer-based connector to align its output with the decoder's text encoder representation, with the same image description. VL-GPT [195] further introduces image semantic loss to enhance multimodal interaction, while Spider [190] explicitly retains the textual output, integrating visual modality features into a constrained textual space. To ensure a lower bound on generation quality, GILL [194] performs image-text retrieval using latent representation and trains a linear classifier to evaluate the quality of generated versus retrieved images.

Pros and Cons. Since the decoder remains frozen during training in these approaches, optimization is relatively stable and easy to converge. However, generation quality is fundamentally constrained by the fixed decoder's capacity, resulting in limited visual fidelity and expressiveness.

Alignment with Image Space. These alignment paradigms [137], [155], [144] first train the decoder to align with the continuous image features output from a frozen vision encoder by treating those features as conditions. During the subsequent joint training with the LLM, optimization is performed using an MSE loss between the latent representation and the vision encoder's image features, while the vision encoder remains frozen. At inference time, the decoder is involved to generate images conditioned on aligned latent representation. Specifically, Emu2 [137], WeGen [199], and Pisces [201] align the decoder with EVA-CLIP [53] features; PUMA [151] and BLIP-3o [23] align the decoder with CLIP [31] features; MetaMorph [155] aligns the decoder with SigLIP [33] features; Seed-X [144] and Nexus-Gen [131] align the decoder with the vision encoder features from Qwen-VL [57] and Qwen2-VL [59], respectively.

To enhance the stochasticity of image feature prediction from the LLM, WeGen [199] adopts in-context learning combined with the sampling procedure to produce diverse self-rewriting captions as intermediate conditions, thereby introducing controlled randomness. Furthermore, BLIP-3o [23] replaces the MSE loss with a flow matching [69] objective to train an additional DiT-based decoder [284], enabling the LLM to predict image features in collaboration with a diffusion model in a more variable manner.

Dynamic Alignment. In methods as UniWorld-v1 [200], MetaQueries [166], OpenUni [173], VisionLLM v2 [198], Ming-Omni [174], *et al.*, the decoder is jointly trained with the LLM in an end-to-end manner, with its parameters being updated during training. Optimization is performed by minimizing the diffusion loss between the generated image and the ground-truth image.

For tasks such as image editing or manipulation, external generation approaches often face a bottleneck in information transmission between the LLM and the decoder, making it difficult to ensure visual consistency between the generated and the input (conditional) image. To mitigate this, several methods introduce visual priors injection, as shown in Fig. 8 (3), and the strategies differ. *1. Latent-Space Injection.* SEED-X [144], CoDi-2 [136], and Ovis-U1 [177] follow the InstructPix2Pix [270] paradigm by concatenating VAE features of the conditional image with the randomly initialized noise of the decoder, thus reinforcing fine-grained visual priors. *2. Condition-Space Injection.* OmniGen-2 [176] feeds the VAE features as the decoder's condition. Differently, X-VILA [205] and UniWorld-v1 [200] leverage semantic features extracted from ImageBind [39] and SigLIP [33], respectively, as decoder conditions. Additionally, MM-Interleaved [139] integrates a deformable attention module [285] within the decoder to interact with multi-scale CLIP [31] features, UniWorld-v1 [200] emphasizes that semantic features are superior to low-frequency VAE features, and further increases loss weights on the edited region. enhancing the model's capacity to incorporate detailed visual information.

To dynamically guide the learning direction of latent representations for varying task requirements, several methods introduce multi-granular tokens. Vitron [191] decouples the latent representation of task-specific decoders into task-specific and task-invariant fine-grained features via adversarial learning [286]. The latter are shared across different tasks to construct a cross-task synergy learning mechanism. Some methods redesign learnable queries to achieve more flexible control: VisionLLM v2 [198] assigns independent learnable queries to task-specific decoders, tailored to the required feature density. Conditional generation tasks receive more queries to capture rich details, while perception tasks are allocated fewer queries focused on essential object-level semantics. Ming-Omni [174] constructs multi-scale learnable queries where image features at different levels directly supervise corresponding latent tokens, enabling explicit semantic alignment across hierarchical levels of learnable queries. As for image alignment setting, PUMA [151] trains multiple SDXL [27], each aligned with CLIP [31] features at a specific granularity. The LLM regresses from coarse to fine levels: similarly, coarse-grained features are used for image generation, while fine-grained for editing, balancing diversity and consistency across tasks.

(2) Video

Due to the use of complete diffusion models as decoders, video-related tasks can be seamlessly addressed by adopting the corresponding video diffusion models.

Text-to-Video Generation. Emu2 [137] adapts the 2D U-Net of SD 2.1 [15] into a 3D architecture by inserting 1D temporal convolutions after each 2D spatial convolution layer, and extending spatial attention into spatio-temporal attention. Several works adopt open-source text-to-video models: Vitron [191], CoDi-2 [136] and NExT-GPT [131] utilize ZeroScope [274], and X-VILA [205] employs VideoCrafter2 [276]. VideoCrafter2 [276] builds on the VideoCrafter1 [287] architecture, fine-tuning spatial modules with a small amount of high-quality synthetic images while preserving its learned motion prior, thereby enhancing overall video quality.

Image-to-Video Generation. Vitron [191] uses I2VGen-XL [275], a cascaded video generation model based on SD 2.1 [15]. In the base stage, high-level semantics are extracted from the input image using a CLIP [31] image encoder to condition the LDM, while low-level details are captured via VQGAN [96] and concatenated in latent space. This design ensures the generated video preserves the input image's structure and intent. The refinement stage applies a separate LDM to enhance resolution, spatiotemporal consistency, and visual sharpness.

(3) Audio

Similar to video modality, off-the-shelf audio diffusion models are employed as decoders.

NExT-GPT [131] and X-VILA [205] utilize AudioLDM [278], which employs a latent diffusion framework for text-to-audio synthesis. AudioLDM transforms raw audio into Mel-spectrograms via STFT and Mel filtering, then compresses them into a low-dimensional latent space using VAE. The diffusion model operates in this latent space for improved efficiency. During training, AudioLDM conditions on CLAP [257] embeddings and trains exclusively on audio data, eliminating dependence on scarce audio-text pairs. At inference, CLAP text encodings guide the diffusion process to generate latent audio features, which are decoded into Mel-spectrograms via VAE and reconstructed into waveforms using HiFi-GAN [288] vocoder.

CoDi-2 [136] and MuMu-Llama [157] adopt AudioLDM 2 [279]. AudioLDM 2 [279] significantly enhances the original UNet-based AudioLDM [278] architecture by integrating a Transformer-UNet backbone, enabling better modeling of long-horizon dependencies and complex structures in audio. Instead of using CLAP embeddings for conditioning, AudioLDM 2 [279] leverages semantically structured representations (LOA, Language of Audio) extracted from AudioMAE [289]. These are injected via cross-attention layers throughout the diffusion process, improving controllability and semantic alignment across tasks such as text-to-audio, speech, and music generation.

5.1.2 *Internal Generation*

(1) Image

Internal generation methods directly inject latent representations into the decoder of diffusion models, *e.g.*, VAE decoder in [15], [103], [14], [277]. In this setting, the generation of latent representations by the LLM is structurally synchronized with the denoising process. During inference, only the key-value cache of the clean image is needed, and CFG can be achieved by simply masking the prefix condition tokens, as illustrated in Fig.9 (1). We categorize methods into ***Latent-level Denoising*** and ***Token-level Denoising*** based on the denoising target of the diffusion model.

Latent-level Denoising. Transfusion [117], LMFusion [231] and MonoFormer [228] perform denoising over all visual tokens directly or incorporate an additional U-Net [102] for upsampling, effectively increasing the resolution of the latent space, as illustrated in 8 (4). The final image is reconstructed using a VAE decoder. JanusFlow [152], X-Fusion [232], BAGEL [22], Mogao [24], and Show-o2 [175] adopt flow matching [69], [245] to formulate a linear denoising objective, which improves training and inference

efficiency. Notably, JanusFlow [152] integrates ConvNeXt blocks [290] and pixel-shuffle layers [291] to further enhance the upsampling ratio, thereby enabling higher latent resolutions. CoDi [128], D-DiT [225], OmniFlow [227], *et al.*, retain the original backbone architectures of diffusion or flow matching models, preserving their decoding processes without modification.

To address the issue of domain gap mentioned above, several methods explore clean-noise decoupling strategies: BAGEL [22] applies teacher forcing to perform complete denoising over all T image tokens in a single training pass, from timestep $t = T$ to $t = 0$. During subsequent generation, the denoised image is used as context, and future tokens are designed to skip over the noisy image tokens via attention masking, attending only to the final clean image tokens. This design explicitly avoids interference from noisy features during training. Furthermore, Mogao [24] arranges the generation task tokens after the understanding task tokens for the same image and performs denoising on them separately. This design mitigates the potential token explosion issue during training caused by complete teacher forcing from BAGEL [22]. In addition, Mogao [24] proposes Multi-modal Classifier-Free Guidance (MCFG), as illustrated in Fig. 9 (3), which decouples the CFG weights of the text and image conditions. By down-weighting the image condition, MCFG reduces over-dependence on conditional image features and enhances the diversity of the generated outputs.

Pros and Cons. These methods allow bi-attention over all visual tokens and enable simultaneous denoising of multiple tokens, facilitating rich visual feature modeling. However, due to the mismatch between noisy tokens used during training and clean tokens as context during inference, a domain gap arises in the decoding phase. It forces the Unified LLM learning from noisy vision features, which may hinder the model's visual understanding and generalization capabilities [22].

Token-level Denoising. Harmon [161] and MMAR [219] combine Masked Autoregressive Reconstruction methods, *e.g.*, MAR [240] and FLUID [292], with the LLM framework. These models adopt bi-attention and predict multiple masked tokens in a random order. An additional MLP-based diffusion module iteratively denoises the predicted tokens and maps them back to the latent space. Once the masked autoregressive process completes, a VAE decoder reconstructs the final image, as illustrated in 8 (5). Furthermore, LatentLM [221], UniFluid [211], and Orthus [220] integrate the masked autoregressive process entirely in the LLM paradigm by employing causal attention to autoregressively generate one token at a time.

Due to the continuous autoregressive plus diffusion paradigm, the inference process suffers from sampling uncertainty and exposure bias. To address this, LatentLM [221] introduces **noise perturbation** to train a σ-VAE. It injects Gaussian noise into the latent space, enabling the VAE decoder to better adapt to input perturbations and thereby improving the stability of autoregressive generation.

Pros and Cons. These methods fundamentally mitigate the domain gap issue inherent in latent-level denoising. While the denoising process is decoupled from the LLM, it requires sequential denoising of each token, thus prohibiting parallel denoising of all tokens.

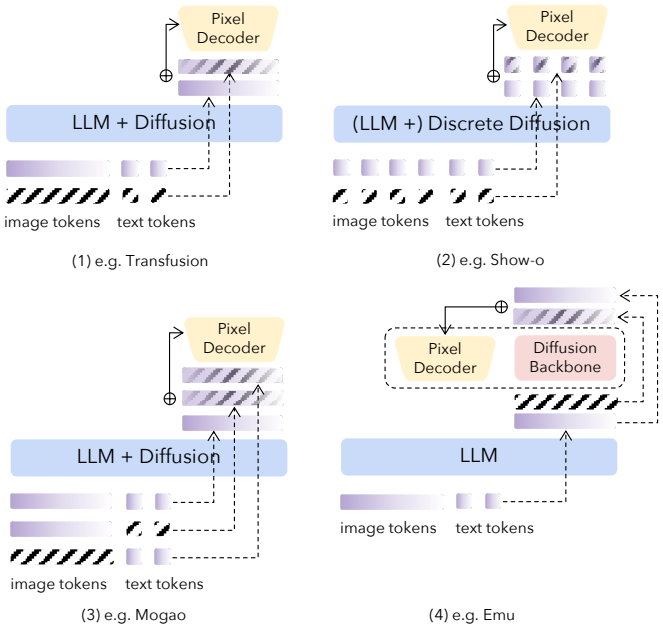

Fig. 9: Typical Classifier-free Diffusion Guidance of UFMs. For diffusion within the UFMs' backbones, CFG is achieved by randomly masking the multimodal instructions, whereas in the extra diffusion modules of UFMs, they are implemented via random dropping of condition embeddings.

(2) Video

Some methods, *e.g.*, LatentLM [221], retain image-based decoding architectures and adopt generate videos by sequentially producing frames in temporal order.

BAGEL [22] improves temporal consistency by randomly grouping video frames and applying bi-attention within each group to enable synchronous denoising across multiple frames. It further utilizes diffusion forcing [293], which perturbs preceding frames with noise to enhance the rollout robustness of long video generation. Spider [190] trains a Mixture-of-Experts (MoE)–based Unified Decoder Connector to align multiple modality-specific decoders, enabling synchronized multi-modal generation. Show-o2 [175] leverages the 3D causal VAE [277] to unify video and image modality processing.

(3) Audio

For audio modality, audio VAE decoders are commonly used to convert latent space features into Mel-spectrograms.

LatentLM [221] trains the proposed σ-VAE on audio data. CoDi [128] adopts the VAE decoder from AudioLDM [278], i.e., HiFi-GEN, to produce Mel-spectrograms, and employs HiFi-GAN [288] as a vocoder to reconstruct waveform signals. OmniFlow further incorporates Rectified Flow [245] to enhance latent modeling, and leverages HiFi-GAN from SpeechT5 [294] as the vocoder to improve the quality and naturalness of audio synthesis.

5.2 Discrete Representation

Decoding strategies for discrete latent representation typically are based on VQ-based VAE decoders of discrete generative models, which take the discrete-space latent features

from LLM and reconstruct back into the continuous target modality data.

Constructing Discrete Latent Representaion. Discrete latent representations are often encoded as discrete tokens. The modality-specific vocabulary is appended to the original text vocabulary, along with modality-specific separators, *e.g.*, <image>,<audio> or task-specific control tokens, *e.g.*, <generation>, <editing>. Lumina-mGPT [148], *et al.*, introduce special tokens to indicate image dimension, *e.g.*, <384>,<512>, and mark the beginning and end of image <SOI>,<EOI>, and the end of image lines <EOL> , enabling more flexible control over image resolutions. OmniMamba [163] adopts separate vocabularies for different modalities and selects the appropriate one based on the task modality, thereby avoiding cross-modal output confusion.

5.2.1 Image

During training, the LLM is optimized with cross-entropy loss between the predicted discrete tokens and the ground-truth image discrete tokens. At inference time, once the full set of image tokens is predicted, each discrete one-hot token is mapped through the VAE decoder's codebook into a quantized feature representation. These quantized features are then fed into the VAE decoder, which reconstructs the corresponding continuous image.

(1) Discrete Auto-Regression.

During training, the LLM is only required to perform the next token prediction based on discrete image tokens, as illustrated in Fig. 8 (6). As for the decoder, typical methods, *e.g.*, CM3 [212], Divter [184], DaVinci [234], OFA [123], RA-CM3 [213], Unified-IO [97], and Unified-IO 2 [138] directly reuse the pretrained vocabulary and VAE decoder from VQ-GAN [96]. To enhance resolution and encoding quality, subsequent works have replaced VQGAN [96] with more advanced image tokenizers. For example, CM3Leon [214] and JAM [98] adopt Make-A-Scene-VQIMG [280]; LWM [142] utilizes aMUSEd-VQGAN [295]; UGen [208] employs SBER-MoVQGAN [296]; while Janus [150], TokLIP [170], and OmniMamba [163], *et al.*, utilize the improved LlamaGen-VQGAN [239]. In parallel, several models such as VARGPT [158] and VARGPTv1.1 [167] adopt the VAR [115] decoder for next-scale prediction, enabling faster generation at higher resolutions. To better align VAE decoders with the requirements of large-scale Unifid LLM, some methods further retrain the tokenizer on larger datasets. For instance, Chameleon [145] retrains Make-A-Scene-VQIMG [280], and Emu3 [25] retrains SBER-MoVQGAN [296] to improve compatibility and performance in downstream tasks.

Token Compression for Efficient Decoding. To improve decoding efficiency and reconstruction resolution, some methods introduce a compression module to reduce token length. For example, Synergen-VL [264] proposes Vision Token Folding and Unfolding, each for increasing feature downsampling and upsampling ratio. The Unfolding is a hierarchical decoding strategy with an additional small causal transformer that progressively reconstructs tokens in multiple stages.

Improving Semantic Fidelity in Discrete Space. To address the loss of high-level semantic information in discrete spaces, several works have proposed novel tokenizer designs that enhance the model's ability to capture semantic features. *1. Implicit Semantic Alignment.* One line of work adopts contrastive learning to implicitly align discrete tokens with semantic representations. VILA-U [149] introduce contrastive loss over quantized tokens to align them with high-level CLIP [31] text semantics. UniToK [159] further incorporates a GAN [96] loss to enhance representation capacity. QLIP [160] leverages a two-stage training strategy, where contrastive learning and adversarial training are decoupled to reduce memory usage and improve training efficiency. *2. Explicit Semantic Disentanglement.* Another direction employs dual-branch architectures to explicitly disentangle the semantic and appearance information. TokenFlow [154] and SemHiTok [216]employ two branches, each corresponding to a separate vocabulary, to reconstruct semantic and pixel information independently. The vocabularies are then concatenated to enforce that the new discrete embedding space carries complete visual information. In contrast, MUSE-VL [215] fuses dual-branch features and shares a discrete embedding for joint semantic and pixel reconstruction, thereby avoiding the need for post-hoc vocabulary operations.

Pros. and Cons. During training, the output of both image and text modalities in the LLM is unified into an auto-regressive format. However, auto-regressively modeling the feature sequence from the VAE image latent space does not align with the intrinsic structure of images, making it less favorable for the LLM to learn. Moreover, auto-regressive decoding of image features is highly inefficient for the LLM, as it typically requires more than $1k$ tokens [25].

(2) Discrete Denoising.

Show-o [116] with followers [229], [171], [230], [226], retrains MAGVIT2-style [282] causal CNN tokenizer to construct a discrete diffusion model via a Masked Token Prediction objective, where visual features are processed through full attention mechanisms in LLM, as illustrated in fig. 8 (7). During training, tokens randomly masked as <MASK> are predicted in parallel. During inference, all tokens are initialized as <MASK> and are gradually predicted step by step, retaining only high-confidence outputs [116], and CFG [283] can be achieved by masking the prefix condition tokens, as illustrated in Fig.2 (2). The final predicted token sequence is decoded into pixel space using the MAGVIT2-style [282] decoder. Additionally, methods *e.g.*, UniDisc [162], UniCMs [226], MMaDA [172], extend this Masked Token Prediction task to the text modality, enabling discrete diffusion across all modalities.

Building on the flexible decoding framework and open-source ecosystem of Show-o [116], several works have proposed further advancements. UniGen [230] enables test-time scaling for Unified LLMs through multimodal self-critique, generating multiple latent token sequences for the same prompt and guiding the LLM to assess image quality step-by-step. UniCMs [226] introduces consistency distillation [101] by obtaining the full denoising trajectory from Show-o's discrete diffusion process, then training the model to directly predict the final output at randomly sampled timesteps, establishing a consistency mapping that accelerates inference. UniCTokens [171] proposes personalized generation through unified concept tokens, where each concept is represented by a set of tokens used across understanding, generation, and unified tasks.

Pros. and Cons. In the discrete space, multiple tokens can be denoised in parallel, yielding an orders-of-magnitude speedup in inference compared to the autoregressive paradigm [116]. However, similar to latent-level denoising, this approach suffers from the mismatch between masked tokens used during training and full tokens as context during inference, leading to a train–inference discrepancy in the image domain and consequently degrading the UFMś capability.

5.2.2 Video

Some works, *e.g.*, VILA-U [149] and LWM [142], directly adopt the image generation paradigm by sequentially generating video frames along the temporal axis. However, this approach significantly increases inference overhead for video generation. To mitigate this, Emu3 [25] incorporates temporal residual layers with 3D convolution kernels into the MoVQGAN [260] architecture and retrains the tokenizer to enhance video tokenization capabilities. By achieving compression across both temporal and spatial dimensions, this design substantially improves the unified model's capacity for generating longer, higher-resolution videos.

5.2.3 Audio

Several works explore the use of audio VQ-based VAE. Unified-IO 2 [138] leverages the existing VIT-VQGAN [297] architecture for training an audio tokenizer with audio datasets. Other methods, *e.g.*, AnyGPT [140] and MIO [202], directly adopt off-the-shelf audio tokenizers like SpeechTokenizer [263] to obtain discrete acoustic units. To further enhance generation granularity, C3LLM [206] utilizes the tokenizer from EnCodec [262] and employs a two-stage decoding strategy: a language model first predicts the coarse-grained tokens (first-layer codes), followed by a non-autoregressive decoder that refines these into fine-grained representations (second-layer codes).

5.3 Hybrid Representation

Hybrid decoding strategies integrate discrete and continuous modeling principles by leveraging LLMs to generate discrete latent representations, which are subsequently transformed into continuous space for diffusion-based generation. A critical component in hybrid decoding is bridging the discrete-continuous latent gap to enable diffusion model compatibility. Methods such as LaVIT [133], Tar [266], and DDT [204] directly utilize pre-trained VQ codebooks, converting one-hot discrete tokens into quantized continuous features through codebook indexing. Alternatively, SEED [130], ILLUME [164], and UniCode² [268] employ trainable code decoders that transform discrete tokens into continuous features without relying on pre-trained codebooks, providing enhanced flexibility and adaptability.

5.3.1 Image

Hybrid representation decoding requires joint training of encoder and decoder components. The decoder either undergoes fine-tuning of the diffusion model or training of a connector module to align discrete tokens with the diffusion process. Following the External Generation taxonomy, methods are categorized by their alignment space. Unlike External Generation, Dynamic Alignment is not applicable here due to the separate decoder training requirement.

Alignment with Text Space. These methods utilize the transferred discrete tokens as input for the diffusion model, which is kept frozen in training, as shown in Fig. 8 (8). SEED [130] with followers [135], [140], [202], trains a reverse Q-Former [196] as a code decoder that transforms discrete causal tokens into continuous features. The diffusion model is frozen and used as the decoder. Only the Q-Former is trained to produce condition features for the diffusion model, optimized through the diffusion loss to align with the model's text space. UniCode² [268] uses a simple MLP as the code decoder, trained with contrastive learning to directly align with the text space of a diffusion model. The diffusion model is not involved in training and is only used during inference.

Alignment with Image Space. Some methods *explicitly reconstruct semantic feature maps from discrete tokens* before aligning them with diffusion models, as shown in Fig. 8 (9). LaVIT [133] first trains a Q-Former [196] to reconstruct ViT-style semantic feature maps from sparse quantized features, and then fine-tunes a diffusion model conditioned on the reconstructed features to align with the image semantic space. Similarly, ILLUME [164] and Tar [266] recover continuous features from complete discrete representations, which are then used to condition diffusion models. Other methods bypass reconstruction in inference and directly use quantized features to condition the diffusion process. ILLUME+ [203] extends this by introducing a dual-branch reconstruction decoder to recover both semantic and pixel-level features. Instead of using the reconstructed results directly, the two branches of quantized features are concatenated and injected into the noise initialization of the diffusion model, merging both high-level and low-level information. FOCUS [265] further incorporates a segmentation model whose predicted masks are used as spatial guidance to the diffusion model, enabling controllable generation and image editing. DDT [204] progressively feeds quantized features into the diffusion model according to the diffusion timestep. This transforms the spatially distributed discrete tokens into a recursive token stream aligned with visual granularity,

Pros and Cons. While this paradigm reduces LLM training burden by relying solely on next-token prediction with decoders active only during inference, it inherits External Generation's information bottleneck, limiting fine-grained detail control. Additionally, independent decoder training precludes joint optimization with LLMs, constraining scalability and model co-adaptation.

5.3.2 Video

Some methods, *e.g.*, MIO, [202], FOCUS [265], adopt a frame-by-frame generation strategy. However, directly scaling LLM's discrete tokens of video results in the tokens number growing linearly with video length, posing significant computational challenges.

To address this, Video-LaViT [141] proposes a Motion-aware Video Decomposition mechanism that factorizes video clips representation into key frame tokens and temporal motion vectors [298], reducing the required token count by up to 90%, as illustrated in Fig. 8 (10). During training, the LLM autoregressively predicts the key frame and motion

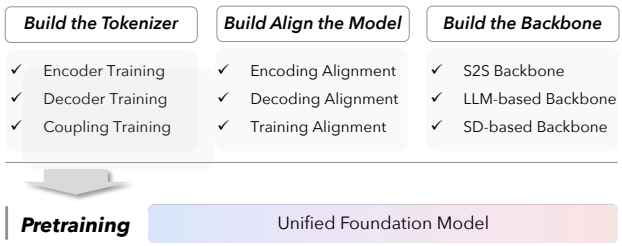

Fig. 10: The pre-training of UFMs. The pre-training process of the model is categorized based on its parameter composition into encoder-decoder, alignment, and backbone modules. This figure illustrates the mainstream methods for building each of these modules.

tokens of the clips. At inference time, it adopts a streaming decoding strategy: first, the image diffusion decoder from LaViT [133] is used to generate the current clip's key frame, conditioned on the key frame token and initialized with the last frame of the previous clip; then, a separately trained video diffusion model with 3D U-Net serves as the video decoder, conditioned on the motion vector and initialized with the key frame to generate the full clip.

5.3.3 Audio

LM-MSN [207] employs a DiT-based [16] model conditioned on discrete tokens, utilizing conditional flow matching for reconstruction training with ultra-low bitrate representation to enhance reconstruction quality. M2-Omni [192] and Ming-Omni [174] leverage pretrained CosyVoice [193] flow matching models, feeding semantic discrete tokens into DiT modules [16] for Mel spectrogram generation. To compensate for semantic-only token encoding, speaker embeddings are integrated to preserve vocal timbre and enhance fidelity.

6 Build the UFM for Pre-training

The pre-training of UFMs aims to construct a single system capable of simultaneously handling multimodal understanding and generation tasks. To systematically analyze its construction paradigm, this section will elaborate from two dimensions: pre-training modules and pre-training strategies. In the section on pre-training modules (Sec. 6.1), we summarize the core parameter composition during the pre-training phase by modularly deconstructing existing models. In the section on pre-training strategies (Sec. 6.2), we systematically review the common methods employed by these models concerning training objectives, data formats, and training procedures. And we can have an overview of training strategies for various models in Tab. 4.

6.1 The modules for pre-training

As illustrated in Fig. 10, the architecture of a UFM can be deconstructed into three core modules: the Encoder-Decoder (Sec. 6.1.1), the Alignment module (Sec. 6.1.2), and the Backbone (Sec. 6.1.3). Each module serves a distinct function and is optimized using specialized pre-training strategies. In this section, we systematically review the construction paradigms for each of these components, detailing their respective architectures and pre-training methodologies.

6.1.1 Encoder-Decoder for Pre-training

The encoder and decoder modules are responsible for the bidirectional conversion of multimodal data between human-perceptible forms and machine-processable internal representations. The encoding (Sec. 4) and decoding (Sec. 5) of UFMs have been explained in detail from the perspective of different modalities previously. Based on the interdependence between the encoder and decoder during the training process, this section divides their construction paradigms into two categories : coupled training and decoupled training. As illustrated in Fig. 11, the following will first discuss the construction processes of the encoder and decoder separately, and then elaborate on the coupled encoder-decoder training paradigm.

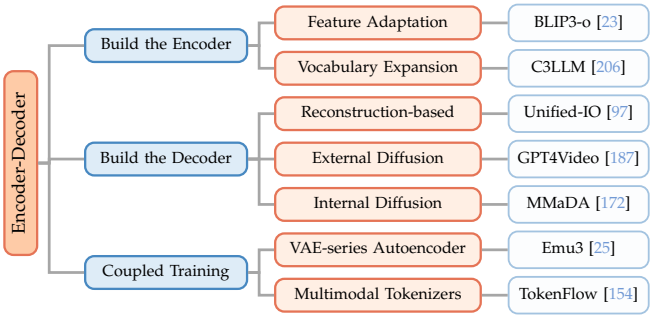

Fig. 11: Build the Encoder-Decoder for pre-training. The construction of encoders and decoders can be either decoupled or coupled, with different construction approaches corresponding to distinct methodologies.

(1) Build the Encoder

The encoder is tasked with transforming raw multimodal data into internal representations suitable for processing by the model's core. Based on their parameterization and training strategies, encoder construction paradigms can be categorized into two principal approaches: a feature adaptation paradigm and a vocabulary expansion paradigm. The former relies on adjusting features from pre-trained extractors, while the latter achieves discrete encoding through vocabulary construction. These two pathways differ significantly in their parameter optimization methods and representational forms.

Feature Adaptation Paradigm. The feature adaptation paradigm primarily relies on pre-trained feature extractors, whose outputs are subsequently refined by a trainable feature adjustment module. This paradigm is characterized by its parametric composition, which includes two core components: a feature extractor with pre-trained, often frozen, parameters, and a feature adjustment module with trainable parameters. The former maps raw multimodal data into a high-dimensional feature space, while the latter aligns these features with the input requirements of the model's core.

For feature extraction, researchers typically employ high-performance pre-trained models to leverage their robust representation capabilities. For instance, the CLIP [31], known for its exceptional image-text alignment, has been utilized as an image feature extractor in works such as FROMAGe [299], GILL [194], and BLIP3-o [23]. Other feature extractor, such as SigLIP [33], have been applied in studies

TABLE 4: An overview of training strategies for various models. We summarize which components are trained during the pre-training stage (Encoder & Decoder, Alignment module, and Backbone) and whether Supervised Fine-Tuning (SFT) and Alignment Fine-Tuning (AFT) are utilized in the fine-tuning process. A checkmark (✓) indicates that a component is trained or a method is used, while a cross (×) indicates the opposite.

Model	Pre-training			Fine-tuning		Model	Pre-training			Fine-tuning	
	Enc. & Dec.	Align	Backbone	SFT	AFT		Enc. & Dec.	Align	Backbone	SFT	AFT
Divter [184]	✓	×	✓	✓	×	VisionLLM v2 [198]	✓	✓	✓	✓	×
CM3 [212]	×	×	✓	✓	×	MoMA [217]	✓	×	✓	×	×
OFA [123]	✓	×	✓	✓	×	TextHarmony [146]	×	×	✓	✓	×
UNIFIED-IO [97]	×	×	✓	✓	×	ANOLE [147]	×	×	×	✓	×
DaVinci [234]	×	×	✓	✓	×	Transfusion [117]	✓	×	✓	×	×
RA-CM3 [213]	×	×	✓	×	×	Show-o [116]	✓	✓	✓	✓	×
FROMAGe [299]	×	✓	×	✓	×	VILA-U [149]	✓	✓	✓	×	×
Visual ChatGPT [125]	×	×	×	×	×	MIO [202]	✓	✓	✓	✓	×
HuggingGPT [126]	×	×	×	×	×	Emu3 [25]	✓	×	✓	✓	✓
AudioGPT [120]	×	×	×	×	×	Janus [150]	✓	✓	✓	✓	×
VALOR [267]	×	✓	✓	✓	×	PUMA [151]	×	×	✓	✓	×
GILL [194]	×	✓	✓	✓	×	JanusFlow [152]	×	✓	✓	✓	×
CoDi [128]	×	✓	✓	✓	×	MoT [153]	×	×	✓	✓	×
LLM-CXR [300]	✓	×	×	✓	×	MUSE-VL [215]	✓	✓	✓	✓	×
MuMu-LLaMA [157]	×	✓	✓	✓	×	Sugar [301]	×	✓	×	✓	×
Emu [129]	×	✓	✓	✓	×	LLaMA-Mesh [302]	×	×	×	✓	×
SEED [130]	✓	✓	✓	✓	×	TokenFlow [154]	✓	✓	✓	✓	×
NExT-GPT [131]	×	✓	×	✓	×	SynerGen-VL [264]	×	✓	✓	✓	×
DREAMLLM [132]	×	✓	✓	✓	×	Vitron [191]	×	✓	✓	✓	×
CM3Leon [214]	✓	✓	✓	✓	×	MetaMorph [155]	×	✓	✓	✓	×
JAM [98]	×	✓	✓	✓	×	ILLUME [164]	✓	✓	✓	✓	×
LaVIT [133]	✓	✓	✓	×	×	LMFusion [231]	✓	✓	×	✓	×
SwitchGPT [183]	×	✓	×	✓	×	Orthus [220]	✓	×	✓	✓	×
Minigpt-5 [134]	✓	✓	✓	✓	×	Liquid [156]	✓	×	✓	✓	×
EasyGen [186]	✓	✓	✓	✓	×	OmniFlow [227]	✓	✓	✓	×	×
SEED-LLaMA [135]	×	✓	×	×	×	LatentLM [221]	✓	×	✓	×	×
CoDi-2 [136]	×	✓	×	✓	×	X-Prompt [218]	×	×	✓	×	×
C3Net [223]	×	✓	✓	✓	×	Janus-Pro [26]	×	✓	✓	✓	×
GPT4Video [187]	×	✓	✓	✓	×	D-DiT [225]	×	×	✓	✓	×
Emu2 [137]	×	×	✓	✓	×	UniCMs [226]	×	×	×	✓	×
Unified-IO 2 [138]	✓	✓	✓	✓	×	UniTok [159]	✓	✓	✓	✓	×
MedXChat [303]	×	✓	✓	✓	×	HealthGPT [304]	×	✓	✓	✓	×
VL-GPT [195]	✓	×	✓	✓	×	UniMoD [305]	×	×	×	✓	×
ModaVerse [188]	×	✓	✓	×	×	QLIP [160]	✓	✓	×	✓	×
MM-Interleaved [139]	×	×	✓	✓	×	ChatVLA [306]	×	✓	✓	✓	×
AnyGPT [140]	×	✓	✓	✓	×	M2-omni [192]	×	✓	✓	✓	✓
LLMBind [189]	✓	✓	×	✓	×	Seed-X [144]	✓	×	✓	✓	×
Video-LaVIT [141]	✓	×	✓	✓	×	GRAPHGPT-O [235]	✓	×	×	✓	×
LWM [142]	×	×	✓	✓	×	ARMOR [209]	×	✓	✓	✓	×
Mini-Gemini [143]	×	✓	✓	✓	×	SemHiTok [216]	✓	×	✓	✓	×
Chameleon [145]	✓	✓	✓	✓	×	WeGen [199]	×	×	✓	✓	×
C3LLM [206]	✓	✓	×	✓	×	DoraCycle [229]	×	×	✓	✓	×
TIGER [185]	×	×	✓	✓	×	BAGEL [22]	×	✓	✓	✓	×
X-VILA [205]	×	✓	✓	✓	×	Make Some Noise [207]	✓	×	✓	✓	×
Harmon [161]	✓	✓	✓	✓	×	UniFluid [211]	✓	×	✓	✓	×
ILLUME+ [203]	✓	✓	✓	✓	×	UGen [208]	✓	×	✓	✓	×
VARGPT [158]	×	✓	×	✓	×	MetaQueries [166]	✓	✓	×	✓	×
UniToken [165]	×	✓	✓	✓	×	OmniMamba [163]	×	✓	✓	✓	×
DDT-LLaMA [204]	✓	×	✓	✓	×	Nexus-Gen [168]	×	✓	✓	✓	×
UniDisc [162]	×	×	✓	×	×	MMAR [219]	×	×	✓	✓	×
VARGPT-v1.1 [167]	×	✓	×	✓	✓	Ming-Omni [174]	✓	✓	✓	✓	×
X-Fuison [232]	×	×	✓	✓	×	Pisces [201]	×	×	✓	✓	×
MonoFormer [228]	×	×	✓	✓	×	Show-o2 [175]	×	✓	×	✓	×
BLIP3-o [23]	×	×	✓	✓	×	UniCode^2 [268]	✓	×	✓	×	×
Mogao [24]	×	×	✓	✓	×	OmniGen2 [176]	✓	×	×	✓	×
Ming-Lite-Uni [169]	✓	✓	×	×	×	FOCUS [265]	✓	✓	✓	✓	×
TokLIP [170]	✓	×	✓	✓	×	UniFork [307]	×	✓	✓	✓	×
UniGen [230]	×	×	✓	✓	✓	ULM-R1 [308]	×	×	✓	✓	×
UniCTokens [171]	✓	×	×	✓	×	Ovis-U1 [177]	✓	✓	✓	✓	×
MMaDA [172]	×	×	✓	✓	✓	Janus-4o [178]	×	×	×	✓	×
OpenUni [173]	✓	✓	×	✓	×	Tar [266]	✓	×	✓	✓	×
X-Omni [179]	×	✓	✓	✓	✓	UniPic [181]	×	✓	✓	✓	✓
Lumina-DiMOO [309]	✓	✓	✓	✓	✓	Emu3.5 [210]	✓	✓	✓	✓	✓

like Janus-Pro [26] and Mogao [24]. To handle modalities like video and audio, methods such as VALOR [267] have incorporated specialized pre-trained modules, including VideoSwin [253] and AST [256]. The parameters of these feature extractors are typically frozen during the training of unified models to preserve their general-purpose capabilities, though this is not a strict requirement.

The feature adjustment module serves as a bridge between the feature extractor and the model's core. Its purpose is to align the extracted high-dimensional features with the semantic space of the core model. The architectural design of this module is diverse. The simplest implementation involves a linear projection layer, as demonstrated in early works like FROMAGe [299] and GILL [194]. To enhance flexibility and representational power, researchers have explored more complex structures. For example, Emu [129] employs a causal Transformer to serialize visual features, LaVIT [133] performs post-processing operations such as merging and selection, and PUMA [151] converts CLIP features into multi-granularity embeddings via multi-layer average pooling. The parameters of these feature adjustment modules are optimized end-to-end during the modal alignment or overall training phase to achieve optimal cross-modal semantic alignment.

Pros and Cons. The feature adaptation paradigm effectively preserves original modal information, while the feature adjustment module enhances feature adaptability and generalization. However, the extracted features can contain a degree of informational redundancy. Compared to discrete encoding with an explicit vocabulary, this paradigm's implicit feature utilization can limit model interpretability.

Vocabulary Expansion Paradigm. The vocabulary expansion paradigm discretizes multimodal information by parametrically extending the vocabulary of a pre-trained model. The core of this method is to map data from different modalities into a shared, discrete symbol space, enabling a language model to process multimodal inputs analogously to its processing of text. This paradigm is primarily realized through two technical pathways.

The first pathway expands the model's vocabulary with new, trainable special tokens, which are parametrically implemented as additional vectors in the word embedding matrix. These tokens serve two primary functions: modality differentiation and task guidance. For modality differentiation, special tokens act as placeholders or identifiers for different data types. For instance, EasyGen [186] uses the `<image>` token as a placeholder that is replaced by image features during processing. For task guidance, tokens are used to convey explicit instructions. LLMBind [189], for example, introduces task-specific tokens like `<gen>` and `<edit>` to direct the model's behavior. Similarly, Unified-IO-2 [138] employs tokens such as `[text]` and `[image]` for modality identification and markers like `[R]` and `[S]` to specify the task paradigm, thereby serving a dual role.

The second pathway aims to construct a unified multimodal vocabulary, fundamentally achieving an isomorphic representation of data from different modalities. This approach involves merging the discrete vocabularies of different modalities into a single, larger, unified vocabulary. For example, Chameleon [145] integrates image tokens with text tokens to build a multimodal image-text vocabulary. Similarly, C3LLM [206] directly extends the language model's vocabulary with audio tokens. By eliminating modality boundaries at the parameter level, this method achieves a more thorough and unified encoding of multimodal data.

Pros and Cons. The vocabulary expansion method, centered on discrete encoding, naturally aligns with the structure of human language, offering good interpretability and modality extensibility. However, the discrete quantization process inevitably involves information compression, and the limited vocabulary size can become a bottleneck for the model's representational capability. Maximizing information retention while maintaining the advantages of discrete encoding remains a key challenge for this approach.

(2) Build the Decoder

The decoder converts the model's internal latent representations into human-perceptible formats, such as images or videos. This process reverses the encoding stage by mapping abstract semantic features to concrete outputs. Based on their underlying mechanisms, decoding methodologies are categorized into three principal paradigms: reconstruction-based decoding, external diffusion, and internal diffusion. Each paradigm offers distinct architectural and implementation trade-offs.

Reconstruction-based Decoding. Reconstruction-based decoding operates on the autoencoder paradigm, learning an inverse mapping from latent features back to the original data. This approach is exemplified by VQ-GAN [96], which was widely adopted in early unified models such as UNIFIED-IO [97] and LLM-CXR [300]. Similarly, other variational autoencoder architectures like RQ-VAE [259] have also been employed as decoding modules [149].

The key characteristic is the coupled training of the encoder and decoder. Both components are jointly optimized to ensure semantic consistency between the encoded features and the reconstructed outputs. The training objective is typically a reconstruction loss, formulated as:

$$\mathcal{L}_{recon} = \mathbb{E}[\|x - \mathcal{D}(\mathcal{E}(x))\|^2], \tag{29}$$

where $\mathcal{E}$ and $\mathcal{D}$ represent the encoder and decoder, respectively. The training of a reconstruction-based decoder is inherently coupled with the encoding process. This specialized, coupled training approach will be elaborated upon in *Sec.(3) Coupled training the encoder and decoder*.

Pros and Cons. The primary strength of reconstruction-based decoding is the high degree of consistency between the encoding and decoding processes, which enables the reconstructed images to align more closely with the details specified in the instructions. However, this method necessitates an additional coupled training stage, thereby increasing overall training complexity. Furthermore, as the training is often performed on datasets of limited scale, the reconstruction quality may not match that of specialized, state-of-the-art generative models (such as Stable Diffusion [15] or DALL·E 3 [310]), which are specifically optimized for high-fidelity image synthesis.

External Diffusion. The exceptional performance of diffusion models in generative tasks has spurred the widespread adoption of decoding strategies that leverage external, pre-trained diffusion models. These methods transform

the model's internal feature representations into human-perceptible data. Based on the content in Sec. 5.1.1, we will introduce two different pathways for implementation: text mediation decoding and feature conditional decoding.

The first pathway is text-mediated decoding, where the model first generates a textual description that subsequently serves as a conditional input to a diffusion model for final data generation. This strategy is employed by models such as MedXChat [303] and GPT4Video [187]. While straightforward to implement, this approach is susceptible to information loss due to the inherent expressive limitations of text. For example, when generating images from text, subtle visual details or spatial relationships present in the original data may be omitted or inaccurately described, resulting in outputs that fail to fully capture the intended content; this limitation is especially evident in tasks requiring precise object localization or fine-grained attribute generation.

The second pathway is feature-conditional decoding, which directly utilizes visual features or latent space representations to condition the diffusion process. For instance, models like Seed-X [144], Emu [129], and BLIP3-o [23] use visual features as conditions, whereas LaVIT [133] and CoDi-2 [136] employ latent space variables. Due to the discrepancy between the model's feature space and the input space of the diffusion model, an adapter module is typically introduced for feature alignment. The training objective for this adapter is often formulated as:

$$\mathcal{L}_{adapter} = \mathbb{E}[\|\mathcal{M}_{diff}(\mathcal{A}(f)) - x\|^2], \tag{30}$$

where $\mathcal{A}$ is the adapter, $\mathcal{M}_{diff}$ is the diffusion model, and f is the input feature.

During training, a common approach is to freeze the parameters of the pre-trained generative model and train only the feature adjustment module. This module learns to map the model's output tokens into the latent space of the generator. NExT-GPT [131], for example, adopts this strategy by training mapping modules to project multimodal tokens into the variable space of frozen generators for data decoding. However, a frozen generative module may not integrate seamlessly with the rest of the model. To achieve superior performance, many models opt to fine-tune the generative module as well. For instance, BLIP3-o [23] not only trains its generative module on specially constructed instruction data but also performs targeted fine-tuning to address specific model deficiencies.

Pros and Cons. The primary advantage of this decoding paradigm is its ability to leverage high-quality, pre-trained generative models, often requiring only lightweight fine-tuning or no additional training at all. This facilitates high-fidelity generation and the integration of advanced generative techniques. However, this approach increases the system's modular complexity. Besides, the semantic alignment between the generated content and the user's instructions can be compromised by the intermediate conversion steps.

Internal Diffusion. The internal diffusion paradigm integrates diffusion-like denoising mechanisms directly into the model's architecture, distinguishing it from approaches that rely on external, pre-trained diffusion models. This method employs an iterative denoising process to reconstruct data from corrupted or masked tokens. Instead of invoking an external generator, the denoising mechanism is an intrinsic component of the model's decoding or backbone structure.

Pioneering works such as Show-o [116] exemplify this strategy at the decoding stage. They employ an iterative denoising process that progressively replaces low-confidence mask tokens over multiple rounds of prediction to generate a complete discrete representation. This representation is then passed to a pixel-level decoder (e.g., MagViT-v2 [282]) to synthesize the final output. While the Show-o series integrates diffusion-based decoding internally, these mechanisms are confined to the decoding module and remain distinct from the primary multimodal response generation process. More recent works, including UniDisc [162] and MMaDA [172], have extended this concept to the model's core backbone. MMaDA, for instance, proposes a unified "Uniformly Random Masking in Iterations" strategy, applying a consistent random masking and iterative denoising procedure to both text and image generation, thereby achieving a more profound architectural unification.

The training objective for such methods typically involves a masked prediction loss, which can be formulated as a denoising objective:

$$\mathcal{L}_{denoise} = \mathbb{E}_{t,\mathcal{M}}[\sum_{i\in\mathcal{M}} -\log p_\theta(x_i|x_{\neg\mathcal{M}}, t)], \tag{31}$$

where $\mathcal{M}$ represents the set of masked token positions, $x_{\neg\mathcal{M}}$ are the unmasked tokens, and t denotes the denoising step. This objective can train the model to recover the original tokens from a corrupted version, enhancing its ability to handle incomplete data and ensuring semantic consistency in the generated outputs.

Pros and Cons. The internal diffusion paradigm effectively integrates the principles of diffusion models into an end-to-end trainable architecture, facilitating a more unified model design. However, this approach can introduce implementation complexity and may face challenges related to training stability, as observed in studies such as UniDisc [162] and MMaDA [172].

(3) Coupled training

Coupled training of encoder-decoder pairs develops unified multimodal tokenizers capable of bidirectional conversion between raw data and latent representations. This approach ensures semantic consistency by jointly optimizing both modules through an autoencoder framework that minimizes reconstruction loss (Eq. 29). These tokenizers are typically pre-trained on large-scale datasets (e.g., ImageNet [311], COCO [312], DataComp-1B [313]) and subsequently frozen during main model training to preserve their specialized encoding-decoding capabilities. This paradigm encompasses two primary architectural categories: VAE-series autoencoders and advanced multimodal tokenizers.

VAE-series Autoencoder. In the early implementation of unified models, some research works directly adopted VAE-series autoencoder generators as image tokenizers to achieve encoding and decoding functionality for image modalities. For example, OFA [123], as an early representative work, employed VQGAN [96] as its image modality tokenizer. This tokenizer was pre-trained on the ImageNet [311], with training objectives encompassing not only reconstruction loss but also adversarial loss mechanisms

specific to VQGAN. In multimodal unified models, the pre-trained VQGAN serves as an encoding structure that maps image data to latent space representations. After sequence-to-sequence processing through the encoder-decoder architecture, the decoding stage utilizes VQGAN's reconstruction capability to restore features to human-readable data formats. Similarly, classic methods such as CM3 [212] and UNIFIED-IO [97] also employed VQGAN for image encoding and decoding processing.

With the continued development of VAE-series models, some research began adopting more advanced model architectures to overcome the performance limitations of classical VQGAN. For instance, a representative example is Show-o [116], which introduced 3D Causal VAE as the image-video modality tokenizer, significantly improving multimodal encoding-decoding performance. In more recent works, Emu3 [25] integrated two temporal residual layers with 3D convolution kernels into both encoder and decoder modules based on the MoVQGAN [296], enhancing video tokenization capabilities. However, these works can still be regarded as minor improvements to VAE-series models, while more complex structural modifications will be elaborated upon in subsequent sections.

Pros and Cons. Classical autoencoder architectures, with their established theoretical foundations and straightforward implementation, provided essential infrastructure for early unified model development. However, their inherent structural simplicity and limited scalability have resulted in diminishing adoption. Contemporary research has pivoted toward sophisticated multimodal tokenizers featuring enhanced architectures and expanded functionality.

Advanced Multimodal Tokenizers. Simple autoencoder architectures often fail to meet complex application requirements, particularly the preservation of causal relationships within encoded sequences. Consequently, researchers have developed sophisticated multimodal tokenizers to achieve superior encoding-decoding performance. These tokenizers comprise specialized modular combinations designed for bidirectional multimodal data transformation. The SEED [130] tokenizer exemplifies this approach and has been widely adopted in subsequent research. It integrates a ViT encoder, causal Q-Former, VQ codebook, MLP, and UNet decoder, enabling effective continuous-to-discrete feature transformation and complete image encoding-decoding workflows. Training employs a two-stage pre-training process on datasets such as CC3M [314] to establish comprehensive encoding-decoding capabilities.

Research in the multimodal tokenizer domain remains highly active, with a continuous stream of innovative designs. For instance, TokenFlow [154] employs a dual-branch architecture for semantic and pixel-level information, enabling inter-branch communication through a shared mapping mechanism to handle image data. LatentLM [221] introduces a σ-VAE that combines a language model loss with a diffusion loss to achieve unified multimodal information processing. These advanced tokenizers, based on coupled encoder-decoder training, adhere to the fundamental autoencoder framework while incorporating refined design optimizations to enhance performance. They are typically pre-trained on representative multimodal datasets such as COCO [312] and ImageNet [311].

Notably, these pre-trained multimodal tokenizers exhibit high reusability and can be integrated as encoding-decoding modules in other models. For example, the SEED tokenizer has been successfully applied in several studies, including AnyGPT [140] and MIO [202].

Pros and Cons. Advanced multimodal tokenizers represent the prevailing trend in coupled encoder-decoder development. Such tokenizers are not only convenient but also align with the principle that a tokenizer should possess both encoding and decoding capabilities. However, they are more complex than traditional autoencoders, and their output quality is often inferior to the specialized generative models.

Notably, discrete encoding necessitates a coupled training for the encoder and decoder. This coupling is crucial because the decoder needs to be trained to interpret the semantic meaning of the discrete codes generated by the encoder, thereby ensuring accurate, semantically-aligned reconstruction. The aforementioned hybrid encoding methods in Sec. 4.3 operate on a similar principle, also requiring coupled encoder-decoder training. In contrast, continuous encoding modules are not subject to this constraint. As they represent information within a continuous feature space, they do not require coupled decoder training to interpret a discrete vocabulary.

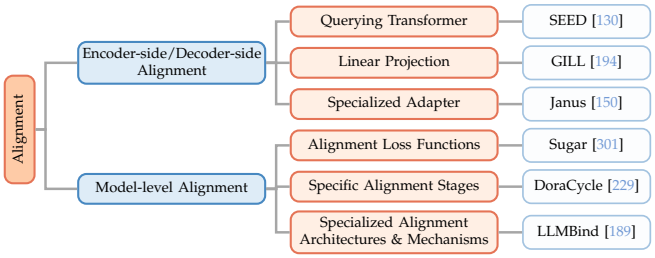

Fig. 12: Alignment during pre-training. Pre-training alignment can occur across different model components, with alignment in distinct components typically corresponding to different parameter structures and methodologies.

6.1.2 *Alignment for Pre-training*

The inherent discrepancies between modalities and distinct representational spaces across modules necessitate alignment strategies during model construction to bridge these gaps. As illustrated in Fig. 12, this section systematically categorizes alignment methods into three primary types: encoding-side alignment, decoding-side alignment, and model-level alignment.

(1) Encoder-side Alignment

Encoder-side alignment modules are designed to map features from various modalities into the representational space of the backbone model. This alignment is crucial for enabling the model to process diverse inputs within a unified framework. Common implementations include linear projection modules and Querying Transformer structures, alongside more specialized alignment architectures, which are also reviewed in this section.

Querying Transformer. Q-former [196] is a lightweight query Transformer architecture that achieves effective cross-modal alignment by enabling learnable query vectors to

interact with visual features. As a classical alignment solution, Q-former serves as a bridge between visual and textual modalities, mapping high-dimensional visual features into semantic representations comprehensible to language models. Its core mechanism lies in utilizing a limited number of query tokens to extract key information from rich visual features, thereby achieving precise cross-modal alignment while maintaining computational efficiency.

The Q-former is a widely adopted alignment mechanism in encoder architectures. For instance, TextHarmony [146] utilizes it to map image features into text-aligned representations, while SEED [130] leverages a similar structure to achieve effective modal alignment and fusion.

Pros and Cons. Q-former effectively extracts salient visual information through limited query tokens and establishes cross-modal alignment between visual and textual representations. However, this method exhibits several limitations. The queries, typically optimized on image-text pairs, have constrained representational capacity that may inadequately capture fine-grained visual details or complex spatial relationships not well-represented in paired text descriptions, leading to suboptimal semantic alignment. Additionally, Q-former incurs higher parameter overhead and training complexity compared to linear projection layers.

Linear Projection. A linear projection layer represents the most straightforward method for feature alignment. This technique has been successfully implemented in several seminal models, such as GILL [194], CoDi [128], and NExT-GPT [131], which utilize linear projection to align multimodal features. These models employ trainable mapping modules to project features from various modalities into the semantic space of a pre-trained language model. By mapping diverse modal inputs into this unified embedding space, the model can achieve effective cross-modal understanding and generation.

While a single linear layer offers a simple alignment mechanism, more sophisticated designs have been developed. For instance, FROMAGe [299] utilizes two distinct sets of mapping layers: one for projecting visual embeddings into the language space and another for mapping both visual and textual features into a shared space for retrieval tasks. Other approaches focus on optimizing the training process. Unified-IO2 [138], for example, enhances its linear projection with a dynamic packing strategy that concatenates multiple training samples into a single sequence, using attention masks to prevent interference. To accommodate multiple input modalities, models like VALOR [267] and X-VILA [205] employ specialized, modality-specific projection modules to facilitate more granular alignment.

Pros and Cons. Projection modules based on linear layers offer simplicity and effectiveness as their greatest advantages. With the continuous emergence of new models in recent years, this method has become the primary choice for alignment in models. However, linear layer structures are simple and can only adjust modal features to a certain extent. Therefore, when facing special alignment requirements, using linear modules alone may be insufficient.

Specialized Adapter. While linear projection and Q-former offer general-purpose alignment, specialized adapters are engineered for scenarios with unique requirements or domain-specific challenges, such as medical imaging or complex multimodal fusion. These adapters provide enhanced flexibility and customization where standard modules may be insufficient. To achieve more refined alignment, researchers have developed advanced modules that move beyond simple linear projection. For instance, EasyGen [186] uses a hybrid MLP and cross-attention adapter for visual features, MedXChat [303] introduces zero-convolution modules for medical images, and VILA-U [149] trains dedicated visual towers with contrastive loss to align discrete visual tokens with textual features. These architectural optimizations enable more sophisticated feature alignment.

Another significant direction is the development of task-oriented dual adapter architectures. Models like Janus and Janus-Pro [26] employ distinct adapters for understanding and generation, enabling targeted optimization for each function. This dual-adapter concept has been widely adopted in subsequent works, including BAGEL [22], Mogao [24], and UniFluid [211], establishing it as a key technical pathway for unified models. Other specialized designs include multi-codebook mechanisms, as seen in TokenFlow [154], which uses a dual-codebook structure to collaboratively align semantic and pixel-level features.

Beyond structural innovations, some studies have explored data-driven alignment methods that rely on specific data formats and interaction strategies. For example, ModaVerse [188] generates "meta-responses" containing invocation instructions for external generative models, aligning language model outputs with generator inputs without requiring additional parameters or training stages. Similarly, VisionLLM-v2 [198] uses hyperlink technology to connect the LLM with task decoders via routing tokens, facilitating effective task alignment. Despite varied implementations, the core objective remains the same: to map encoded modal information into the semantic space of a pre-trained model for unified processing.

Pros and Cons. Specialized adapters can be tailored to specific requirements, yielding more precise modal alignment. However, their development is often more resource-intensive, and their high degree of specialization can limit their generalizability and widespread adoption.

(2) Decoder-side Alignment

During the decoding phase of multimodal models, the backbone's output, typically consisting of abstract feature representations or discrete tokens, needs to be transformed by a decoder into human-perceptible data formats. A semantic gap often exists between the model's output space and the decoder's input space. To bridge this gap, alignment modules are introduced at the decoding stage to ensure effective feature mapping.

The design of decoding-side alignment modules is technically analogous to that of their encoding-side counterparts. Their primary objective is to establish semantic consistency between the model's output features and the input space of the designated decoder. Common implementations include linear layers and attention mechanisms, with the specific choice contingent on the decoder's architecture and task requirements. For instance, when a diffusion model serves as the decoder, the alignment module can map the backbone's output into the diffusion model's condi-

tional space. Conversely, for autoregressive decoders, it is paramount to align the output features with the decoder's embedding space. As the implementation are largely consistent with those used for encoding-side alignment, further details can be found in *Sec.(1) Encoder-side Alignment*.

In terms of training, decoding-side alignment modules are typically optimized jointly with their encoding-side counterparts. Parameters are updated synchronously through an end-to-end training process, often employing a unified loss function and backpropagation. This collaborative optimization strategy simplifies the training pipeline and enforces semantic consistency across the entire encoding-decoding chain, thereby enhancing the model's overall performance.

(3) Model-level Alignment

Beyond discrete alignment modules at the encoding or decoding stages, model-level alignment achieves cross-modal coherence through specialized loss functions and training strategies integrated into the overall optimization process. This approach can be implemented through specific objectives that operate throughout training or through carefully curated data and targeted training procedures. Notably, model-level alignment is complementary to dedicated alignment modules and can be employed concurrently.

Alignment Loss Functions. Contrastive loss functions, such as the one in Eq. 1, represent the predominant approach for multimodal alignment training. Following the breakthrough success of the CLIP in multimodal alignment, contrastive loss has become a cornerstone technology in the field and plays an integral role in training unified models. For instance, C3Net [223] employs contrastive learning to achieve effective alignment across image, audio, and text modalities. This technique is now widely adopted in numerous unified models to enhance cross-modal semantic consistency.

To address specific application requirements, researchers have developed various innovative alignment loss. Minigpt-5 [134], for example, introduces a Voken Alignment Loss designed to ensure precise consistency between generated visual tokens and the conditions of the diffusion model. VALOR [267] proposes a Multimodal Grouped Alignment loss that uses contrastive learning to map text, visual, audio, and audiovisual modalities into a shared semantic space. Similarly, Sugar [301] implements a dynamic sequence alignment loss based on sequence similarity computation. These specialized loss functions are goal-oriented, designed to improve the representational quality and alignment precision of tokens in specific contexts.

Pros and Cons. Alignment based on loss functions seamlessly integrate the modal alignment process into the training pipeline, enabling end-to-end optimization. However, a significant challenge with this approach is that the configuration of loss weights relies on extensive hyperparameter tuning. Balancing different loss functions to achieve optimal results often requires numerous experiments.

Specific Alignment Stages. Beyond alignment-focused loss functions, dedicated training stages enable more precise multimodal alignment through decoupled optimization. This approach separates alignment from the main training pipeline, creating an independent phase that offers targeted control for fine-grained or domain-specific alignment requirements. For instance, MIO [202] implements a dedicated alignment stage using image-text pairs with contrastive and reconstruction losses to map multimodal features into the textual semantic space. Similarly, SynerGen-VL [264] introduces a visual alignment stage to harmonize visual representations with the language model's semantic space, establishing a foundation for subsequent unified training.

Instruction-driven alignment represents another prominent approach. In contrast to traditional pair-based methods, this technique leverages instruction-formatted data to cultivate the model's alignment capabilities. This is exemplified by the Modality-aligned Instruction Tuning proposed in SwitchGPT [183], which first trains the model to comprehend and generate modality alignment instructions and subsequently uses these instructions to guide the generation of more precise cross-modal responses.

Cyclical iterative alignment has also been explored, albeit less commonly. DoraCycle [229] employs an innovative "A→B→A" training path to achieve a comprehensive alignment of semantic relationships between modalities, aiming for modality-equivalent transformations. Although theoretically compelling, experimental results have indicated limited effectiveness, suggesting that the alignment achieved during conventional training may already be adequate for subsequent tasks.

Pros and Cons. Specialized alignment stages facilitate precise capability enhancements and offer fine-grained control over multimodal alignment. However, this strategy increases both training complexity and computational overhead, necessitating a trade-off between performance gains and resource expenditure.

Specialized Alignment Architectures and Mechanisms. Certain methods achieve alignment by designing specialized modular structures and mechanisms that are integrated directly into the model's training process. For instance, LLMBind [189] incorporate Mixture-of-Experts (MoE) [315], [316] modules into LLMs to facilitate modal alignment and dynamically route computations for different task functionalities. LMFusion [231] extends an LLM with a parallel image module, integrating the two through a cross-modal self-attention mechanism. As the text component remains frozen, this cross-modal attention mechanism effectively functions as an alignment module. Similarly, Sugar [301] employs a dynamic sequence alignment mechanism that computes modal similarity via global alignment kernels to achieve semantic alignment.

Pros and Cons. In contrast to employing dedicated modules at the encoding or decoding stages, these methods embed the alignment process directly into the training pipeline. This integration is accomplished by either jointly optimizing the alignment components with the backbone or by incorporating an alignment-specific loss into the overall training objective, thereby embedding alignment within the end-to-end optimization process. However, as these mechanisms are typically tailored to specific requirements, their generalizability is often limited.

6.1.3 Build the Backbone for Pre-training

The backbone is the core component of a unified model, responsible for multimodal semantic understanding and

task response generation. Typically built upon a LLM or a MLLM, the backbone processes encoded and aligned multimodal features to produce the desired output. Backbone construction strategies have evolved over time, leading to three primary approaches: early exploratory models, LLM-based backbones, and diffusion-based backbones, which can be seen in Fig. 13. This section reviews these methods, highlighting their chronological development and fundamental architectural differences.

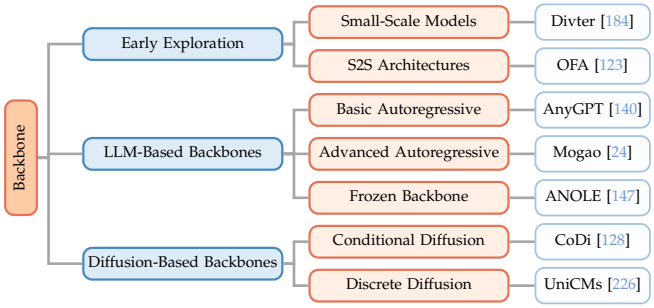

Fig. 13: Build the Backbone for pre-training. The evolution of the UFM's backbone can be summarized along three main lines: early exploratory approaches, methods based on LLMs, and diffusion-based methods.

(1) Early Exploration

Before introducing the mainstream backbone in detail, some early exploratory work is first presented.

Early Small-Scale Models. The unification of understanding and generation tasks has been a long-standing objective in machine learning, with initial explorations predating the era of large-scale pre-trained models. However, these early efforts were typically limited by technical compromises tailored to specific, often narrow, application scenarios.

Divter [184], about the earliest models capable of unified understanding and generation, exemplifies early architectural approaches through specialized module collaboration. The system comprises three components: a text dialogue generator producing multimodal tokens, an image tokenizer for bidirectional image encoding/decoding, and a text-to-image translator mapping image tokens to the tokenizer's latent space. Joint optimization via composite loss functions ensures coordinated training across modules. However, experimental validation was limited to the PhotoChat dataset, restricting demonstrated task coverage.

Pros and Cons. Early small-scale models represented initial attempts at unifying understanding and generation but were constrained by task-specific architectures and functional compromises that limited true generalization. These systems, exemplified by models like Divter with their separate specialized modules for text and image processing, lacked the scalability and flexibility required for complex, open-domain applications.

Sequence-to-Sequence Architectures. The advent of Transformer architectures spurred the development of early sequence-to-sequence (S2S) models, which laid a foundational groundwork for unifying understanding and generation tasks. These models, built upon encoder-decoder frameworks, enabled the unified processing of multimodal tasks. This architectural choice represents a significant departure from the decoder-only architectures that characterize contemporary multimodal large language models.

Representative of these early unified models, UNIFIED-IO [97] was built upon the T5 [317] encoder-decoder architecture. It was pre-trained in an unsupervised manner on large-scale datasets spanning over 90 distinct tasks, with training data encompassing multiple modalities such as text, images, and image-text pairs. Its successor, Unified-IO2 [138], continued this encoder-decoder design philosophy while introducing innovations in training objectives, employing a multimodal mixture-of-denoisers strategy that combines masked denoising with causal generation. Similarly, the models like OFA [123] also leveraged encoder-decoder architectures, developing the model's capacity to execute diverse tasks through multi-task pre-training.

Pros and Cons. Sequence-to-sequence architectures enabled preliminary unification of understanding and generation through large-scale pre-training. However, their substantial training costs and limited performance led to their gradual displacement by autoregressive paradigms, which now dominate contemporary unified model development alongside a smaller body of diffusion-based approaches.

(2) LLM-Based Backbones

The breakthrough advancements of LLMs have introduced a new paradigm for unified multimodal modeling. By leveraging the powerful sequence modeling and semantic understanding capabilities of pre-trained LLMs as the core backbone, researchers have developed a multitude of unified models. This approach typically employs an autoregressive framework, where data from different modalities are uniformly converted into serialized tokens and trained using the NTP objective. This design simplifies the architecture and training pipeline compared to traditional encoder-decoder structures. While the core concept remains consistent, various technical pathways have been explored in terms of implementation details and architectural design.

Basic Autoregressive Paradigm. The basic autoregressive paradigm represents a direct and widely adopted approach for constructing UFMs. Its core principle is to encode heterogeneous multimodal data into a unified sequence of tokens, which are then modeled autoregressively using the classic NTP objective. The training objective for this paradigm is the standard autoregressive loss, formally defined as:

$$\mathcal{L}_{NTP} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}), \tag{32}$$

where x_t is the t-th token in the sequence and $x_{<t}$ denotes its preceding context. This function trains the model to predict the next token based on the preceding sequence, enabling unified modeling of multimodal data.

In terms of implementation, mainstream methods typically initialize their parameters from pre-trained language models to leverage their powerful sequence modeling capabilities. For example, Emu [129] is initialized with the LLaMA model and undergoes unified autoregressive pre-training on multimodal token sequences, using a cross-entropy loss for text and an L2 regression loss for visual components. Similarly, LaVIT [133] performs unified NTP training on multimodal token sequences, and AnyGPT [140] conducts NTP pre-training on text-centric aligned multi-

modal data. These methods exemplify the classic paradigm of initializing from a pre-trained language model.

Some studies have explored alternative approaches that do not rely on language model initialization. For instance, Transfusion [117] builds an autoregressive Transformer model from scratch, CM3 [212] employs a Causally Masked Objective to guide training, and Sugar [301] constructs a unified architecture based on the understanding model VILA [149]. Unlike methods that leverage pre-trained language models, these approaches start from randomly initialized parameters and are trained entirely on multimodal data. Despite differences in initialization strategies, their core architecture remains a direct stack of Transformer layers that model multimodal tokens autoregressively.

The basic autoregressive paradigm has demonstrated enduring vitality and continues to be widely adopted in recent works such as UniTok [159] and DDT-LLaMA [204], underscoring its effectiveness and practicality in unified multimodal modeling. For example, UniTok achieves strong performance in cross-modal retrieval and captioning, while DDT-LLaMA shows notable improvements in multimodal reasoning and instruction-following, further validating the paradigm's advantages.

Pros and Cons. The basic autoregressive architecture is simple and effective, facilitating end-to-end optimization with relatively stable scaling properties, which has led to its extensive adoption in LLMs and MLLMs. However, this approach incurs substantial training costs and is susceptible to error accumulation, making it challenging to model complex pixel-level dependencies.

Advanced Autoregressive Paradigms. While the basic autoregressive paradigm has proven effective, researchers have continuously explored architectural and strategic optimizations to overcome its performance bottlenecks. These advancements can be broadly categorized into two dimensions: architectural enhancements, which involve integrating specialized components, and strategic innovations, which focus on designing composite training objectives.

Architectural optimizations aim to enhance multimodal processing capabilities or improve training efficiency by introducing specialized structures. To bolster understanding capabilities, JAM [98] extends CM3Leon [214] with a new text-image branch, achieving effective fusion via cross-attention mechanisms. LLMBind [189] embeds a MoE mechanism into a Vicuna-initialized LLM, enhancing performance through dynamic expert routing. Mogao [24] employs MIMO Transformer Blocks, integrating multimodal multi-head attention within the Transformer layers to optimize cross-modal interactions. To improve training efficiency, methods such as MoT [153] decouple non-embedding parameters (e.g., feed-forward networks and attention matrices), enabling modality-specific processing with global attention and significantly improving training efficiency across various tasks.

Strategic innovations in training involve designing diverse composite loss functions to enhance training effectiveness while retaining the core autoregressive framework. For instance, MM-Interleaved [139] employs a dual training objective of Next Token Prediction and Next Image Prediction. Minigpt-5 [134] combines a text autoregressive loss, a latent space diffusion loss, and an auxiliary MSE loss for multi-objective pre-training. UniMoD [305] introduces a Masked Token Prediction loss to strengthen model training, and Sugar [301] innovatively integrates generative and discriminative losses to achieve synergistic optimization.

Pros and Cons. Advanced autoregressive paradigms enhance unified model performance through architectural innovations and strategic optimizations, establishing critical technical pathways for the field. However, specialized designs may compromise generalizability, limiting universal applicability across diverse scenarios.

Frozen Backbone. A significant line of research has focused on constructing unified models by freezing the parameters of a large pre-trained backbone and exclusively training lightweight auxiliary modules. This paradigm, validated by early works such as FROMAGe [299] and GILL [194], operates on the principle of projecting diverse modal information into the fixed semantic space of a pre-trained LLM. Within this latent space, the frozen LLM performs cross-modal understanding and task reasoning, and its outputs are subsequently decoded into human-perceptible formats.

This strategy is typically implemented by training lightweight adapter modules for modal alignment. For instance, FROMAGe [299] trains linear layers to map visual features into the language space, while GILL [194] employs specialized adapter architectures to achieve unified image understanding and generation with frozen LLMs. Recent approaches like MetaQueries [166] extend this concept by freezing MLLM backbones and training only learnable queries to condition generation. This design philosophy significantly reduces computational overhead while leveraging pre-trained model capabilities.

As the performance of unified models improves, some studies have begun to explore the UFM based on a frozen unified model backbone. For example, UniCTokens [171] utilizes a frozen pre-trained Show-o [116] model, learning unified concept tokens through a multi-stage training process. Similarly, the ANOLE [147] method freezes the majority of the Chameleon's parameters [145], fine-tuning only a small subset on curated data. However, the generalization capability of the frozen-backbone strategy has inherent limitations, which has restricted its adoption in the development of unified models. Nevertheless, as unified model performance continues to strengthen, this technical approach holds considerable research potential.

Pros and Cons. The frozen-backbone strategy significantly streamlines the UFM's training process, enabling researchers to meet fundamental functional requirements at a reduced computational cost. This is particularly advantageous in resource-constrained environments. However, constructing unified models typically requires coordinating understanding and generation capabilities across diverse tasks, which poses a fundamental challenge for backbones pre-trained on specific tasks. As unified models advance, future developments may yield more sophisticated backbones that can effectively serve this role while remaining frozen.

(3) Diffusion-Based Backbones

While diffusion models excel at generative tasks, their inherent limitations in understanding and reasoning constrain their use as backbones for unified models, which require

strong comprehension to execute diverse tasks. Despite this, researchers continue to explore diffusion-centric paradigms, driven by their generative prowess. This line of inquiry is valuable for advancing technical diversity and establishing empirical foundations for the potential and boundaries of diffusion models in unified modeling.

Conditional Diffusion Models. Conditional diffusion models extend the standard diffusion framework by incorporating external information to guide the generation process. This conditioning, often derived from other modalities, enables more controlled and targeted outputs. In the context of unified modeling, their core principle is to leverage data from one or more source modalities as conditional inputs to steer the diffusion process toward generating a desired target modality output, thereby facilitating cross-modal generative transformations.

CoDi [128] is a representative work that employs a multi-stage training strategy to construct a unified multimodal diffusion architecture. It features a unified conditioning mechanism that maps features from various modalities (e.g., text, image, audio) into a shared conditional space. Subsequently, modality-specific diffusion decoders perform the final generation. This design enables the model to handle generative transformation tasks across multiple modalities within a single diffusion framework. However, while conditional diffusion models are frequently used as decoders, their inherent functional limitations have restricted their adoption as backbone modules, where they have served primarily as a source of exploratory inspiration.

Pros and Cons. Conditional diffusion models excel in high-fidelity multimodal generation through their natural conditioning framework for cross-modal synthesis. However, their generation-centric design inherently limits their understanding capabilities, constraining their effectiveness in complex reasoning and semantic comprehension tasks required for unified model backbones.

Discrete Diffusion Models. Beyond conventional diffusion models, discrete diffusion methods offer a novel paradigm for unified modeling by extending denoising principles to discrete variable spaces. This approach achieves unified generative modeling through diffusion processing of discrete tokens and demonstrates superior adaptability when handling inherently discrete modalities such as text. The training objective can be formalized as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),q(x_t|x)}\left[\frac{\alpha_t}{1-\alpha_t}\log p_\theta(x_0 \mid x_t)\right], \quad (33)$$

where t represents the diffusion timestep, x_t denotes the noisy state at timestep t, and α_t is a hyperparameter.

In specific implementations of unified modeling, UniDisc [162] pioneered the extension of discrete diffusion models to joint text-image modeling, achieving unified generation and understanding across modalities through a designed "mask-unmask" discrete noise process. The core innovation lies in uniformly mapping data from different modalities to discrete token spaces, followed by processing using identical diffusion denoising mechanisms.

Building upon similar concepts to UniDisc, the recent MMaDA [172] further employs a unified MTP objective:

$$\mathcal{L}_{\text{uni-df}} = -\mathbb{E}_{t,x_0,x_t}\left[\frac{1}{t}\sum_{i=1}^{L}\mathbb{I}[x_t^i = \text{[MASK]}]\log p_\theta(x_0^i \mid x_t)\right]. \quad (34)$$

This loss function effectively unifies training objectives for text and images through a unified masking prediction mechanism, where $\mathbb{I}[\cdot]$ is an indicator function and L represents sequence length. This "mask-and-repair" training paradigm provides a novel theoretical framework for multimodal unified modeling. Here, Next Patch Prediction is used for visual modalities, analogous to NTP for text.

In contrast, Show-o [116] was among the earlier methods to propose discrete diffusion for image response generation, though its text component still employs traditional autoregressive generation mechanisms while the image component undergoes discrete diffusion decoding after full attention processing. This design maintains the established advantages of text generation while exploring new paradigmatic approaches to model architecture. Similarly, subsequent works such as UniCMs [226] and Show-o2 [175] have implemented unified model construction based on similar architectural concepts, demonstrating the continued development potential of this technical route, although these methods ultimately fail to achieve truly unified modeling of text and image tokens.

It should be noted that, as demonstrated by UniDisc [162] and corroborated by existing research and experiments, while diffusion algorithms exhibit clear advantages in processing modalities such as images, they still face unavoidable disadvantages in training efficiency for text tasks. Therefore, although this approach of uniformly applying diffusion denoising to both text and images successfully unifies text and image responses within the diffusion framework, the training efficiency of such models still requires further optimization.

Pros and Cons. The discrete diffusion paradigm introduces a novel approach to unified image and text generation by extending the denoising principle to discrete spaces. This method theoretically enables the isomorphic processing of images and text within a unified Transformer architecture, inheriting the high-fidelity generation capabilities of diffusion models. However, its generation-centric design imposes structural limitations on tasks that demand complex semantic understanding and reasoning. Furthermore, the iterative denoising process substantially increases inference latency, thereby compromising overall model efficiency.

6.2 Pre-training Strategies

Pre-training UFMs requires coordinated optimization of multiple components, including the unified tokenizers, the alignment modules, and the backbone networks. The training strategy should ensure these components work synergistically to establish unified multimodal understanding and generation capabilities. This section systematically examines pre-training methodologies across three dimensions: training objectives (Sec. 6.2.1), data formats (Sec. 6.2.2), and training procedures (Sec. 6.2.3).

6.2.1 Training Objectives

The pre-training of UFMs involves the coordinated optimization of multiple components, achieved through a composite loss. Based on the modular framework outlined above, the overall training objective can be formalized as:

$$\mathcal{L}_{unified} = \alpha_1\mathcal{L}_{tokenizer} + \alpha_2\mathcal{L}_{align} + \alpha_3\mathcal{L}_{backbone}, \quad (35)$$

where $\mathcal{L}_{tokenizer}$ is the loss for the encoder-decoder modules, responsible for the effective encoding and reconstruction of multimodal data; $\mathcal{L}_{align}$ is the modal alignment loss, which ensures that features from different modalities are mapped into a unified semantic space; and $\mathcal{L}_{backbone}$ is the loss for the model's core, reflecting its fundamental modeling paradigm (e.g., the Next Token Prediction loss for autoregressive models or a denoising loss for diffusion models). The coefficients $\alpha_1, \alpha_2, \alpha_3$ are weighting factors in the range $[0, 1]$ that balance the contribution of each component. A coefficient is set to 0 if the corresponding module's parameters are frozen, and to 1 to assign maximum importance to its training.

It is important to note that not all unified models incorporate all three loss components simultaneously. In practice, the model backbone is rarely trained independently; instead, it is jointly optimized with other modules, making $\mathcal{L}_{backbone}$ the primary driver of the training process. For models that utilize pre-trained tokenizers or frozen backbone modules, the corresponding loss weights (α_1 or α_3) are set to zero, excluding those components from parameter updates. This flexible loss function design allows different unified models to adopt tailored training strategies that suit their specific architectures and technical approaches while maintaining a consistent optimization framework.

6.2.2 Data Formats

Traditional pre-training on image-text pairs is increasingly insufficient for the complex requirements of unified multimodal models. To achieve more precise cross-modal alignment, researchers have widely adopted instruction-based data construction strategies. These strategies convert heterogeneous modal data into unified sequential representations using structured templates. This approach involves pairing multimodal data (e.g., images, text, audio) with explicit instructions or prompts that specify the desired task or response, such as "`Describe the objects in the image: <image>`" or "`Replace the grass with sand in the photo: <image>`". Unlike traditional image-text pairs, which typically associate an image with a simple caption, instruction-based data provides richer context and explicit task guidance.

For data format standardization, mainstream multimodal instruction data employs a standardized delimiter system. A common format is: *<text>text_content</text> <image>image_tokens</image> <video>video_tokens</video> <audio>audio_tokens</audio>*, where *image_tokens*, *video_tokens*, and *audio_tokens* represent the tokenized forms of their respective modalities (e.g., discrete code sequences for images and videos, or spectrogram/waveform tokens for audio).

In specific implementations, delimiter designs may vary. For instance, specialized tokens like *<imgbos>* and *<imgend>* are sometimes introduced to explicitly denote the beginning and end of image token sequences within the multimodal data stream. These tokens help the model accurately identify image boundaries, facilitating tasks such as image generation, editing, or segmentation, thereby better adapting to the requirements of special models. This special tokens system provides a crucial technical foundation for the end-to-end modeling of multimodal sequences.

Regarding the construction of instruction templates, researchers have designed strategies tailored to different task types. For object detection, a training sample can be created by appending a task-oriented prompt like "What objects are in the image?" to an input image. This template-based method maintains the clarity of task semantics and enhances the model's adaptability to diverse instructions. For generation tasks, instruction data is often created by converting classic image-text pair datasets (e.g., ImageNet, COCO). A typical generation instruction template might take the form: "`What is the complete image? Text: <image caption>`".

To lower research barriers and promote development in the field, the academic community has released several high-quality instruction datasets, such as LLaVA-1.5 [56] and SEED-Data-Edit [318], which provide vital data support for training unified models. LLaVA-1.5 is characterized by its diverse multimodal instruction-following data, making it suitable for tasks requiring visual understanding and dialogue. In contrast, SEED-Data-Edit focuses on image editing instructions, supporting scenarios that involve visual generation and manipulation. The availability of these standardized datasets has significantly accelerated the development and application of unified multimodal modeling. The following examples from LLaVA-1.5 and SEED-Data-Edit illustrate the data formats for understanding and generation instructions, respectively.

```
# understanding example from LLaVA-1.5
{
  "id": "000000442786",
  "image": "000000442786.jpg",
  "conversations": [
    {
    "from": "human",
    "value": "What do you see happening in this
    image?\n<image>"
    },
    {
    "from": "gpt",
    "value": "The scene depicts a lively plaza area
    with several people walking and enjoying their
    time. A man is standing in the plaza with his
    legs crossed, holding a kite in his hand. The
    kite has multiple sections attached to it,
    spread out in various directions as if ready for
     flight.\n\nNumerous people are scattered
    throughout the plaza, walking and interacting
    with others. Some of these individuals are
    carrying handbags, and others have backpacks.
    The image captures the casual, social atmosphere
     of a bustling plaza on a nice day."
    }
  ]
},

# generation example from SEED-Data-Edit
{
   "source_caption": "a black and white photo of a
    young man sitting in the grass",
 "target_image": "unsplash_images/photo
    -1615489843990-6a10170d1470_2.jpg",
 "target_caption": "a black and white photo of a
    young man sitting in the sand",
 "match_score": 0.9678,
 "source_image": "unsplash_images/photo
    -1615489843990-6a10170d1470.jpg",
 "edited object": "grass",
 "instruction": "replace the grass with sand",
 "task_type": "replace object",
},
```

6.2.3 Staged Training

The pre-training of unified models, given the heterogeneity of multimodal data and the complexity of unified tasks, typically follows a staged paradigm to ensure stability and effectiveness. Building on the modular framework previously established, this process can be deconstructed into three principal stages: (1) pre-training of the encoder-decoder module (Sec. 6.1.1), (2) multimodal alignment training (Sec. 6.1.2), and (3) unified backbone training (Sec. 6.1.3).

First, the encoder-decoder pre-training stage equips the model with foundational modal transformation capabilities, typically through objectives like autoencoding or conditional generation, to achieve robust encoding and decoding performance. Second, the multimodal alignment stage is dedicated to bridging the semantic gap between different modalities by mapping their respective features into a shared representational space. Finally, the unified backbone training stage integrates these components through joint optimization, endowing the model with the capacity for both multimodal understanding and generation.

It is important to note that not all unified models undergo all three training stages, nor are these stages always distinct and independent. The specific configuration depends on the model's architecture and technical approach. For instance, some models may merge or omit certain stages. Models like UNIFIED-IO [97] and Janus-Pro [26] bypass a dedicated encoder-decoder training stage by adopting pre-trained components. Similarly, methods such as OFA [123] and D-DiT [225] achieve modal alignment via end-to-end joint optimization, thereby merging the alignment stage with the backbone training rather than treating it as a separate step. Models employing a frozen backbone strategy further simplify the process by focusing primarily on training lightweight adapters.

Furthermore, some UFMs incorporate MoE mechanisms to enhance task adaptation. Within our analytical framework, the training of MoE modules can be viewed as an extension of the multimodal alignment stage. For instance, Uni-MoE [319] integrates a sparse MoE architecture into the LLM backbone. The specialized training of these MoE modules is usually conducted after the initial multimodal alignment and before the final unified training, consistent with the staged alignment paradigm described herein.

As formulated in Sec. 2.3, the objective of unified pre-training is to develop a model equipped with both comprehension and generation capabilities. While the subsequent fine-tuning can further improve model performance, it is not essential to the core training pipeline. In fact, to ensure evaluative fairness and procedural integrity, some studies directly assess the pre-trained model as the final unified system. For instance, models like UNIFIED-IO [97] and UNIFIED-IO2 [138] are explicitly designed for zero-shot unified tasks, and their development is confined to a pre-training phase. Similarly, the sequential training strategy of BLIP3-o [23], which involves distinct stages for developing understanding and generation capabilities, is classified herein as part of the unified pre-training, since a model is not considered unified until it acquires generative abilities. In contrast, the distinction between pre-training and fine-tuning is more ambiguous in studies such as RA-CM3 [213], VILA-U [149], and OmniFlow [227]. Given the ambiguity, this paper establishes a clear demarcation between these two phases to provide a distinct conceptual framework and mitigate potential confusion in future research.

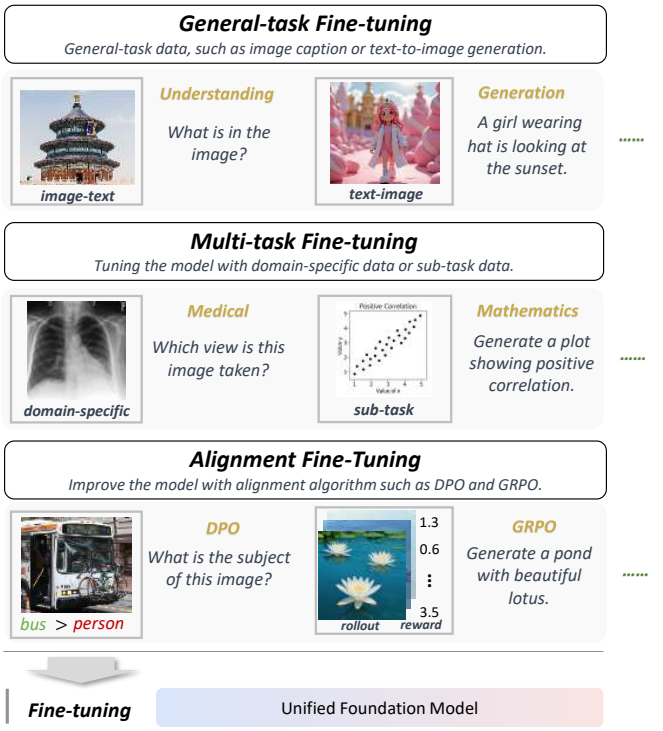

Fig. 14: An overview of fine-tuning for unified models. The approaches are categorized into task-supervised fine-tuning (which includes general-task and multi-task fine-tuning) and human preference-based alignment fine-tuning.

7 Improve the UFM by Fine-tuning

To further enhance the performance of unified models, various fine-tuning strategies have been developed. As illustrated in Fig. 14, these strategies are broadly categorized into two main paradigms: task-supervised fine-tuning (Sec. 7.1) and human preference-based alignment fine-tuning (Sec. 7.2). The former optimizes model performance by learning from annotated datasets without direct human intervention. In contrast, the latter integrates human feedback signals to guide the optimization process. This alignment-based approach has demonstrated significant performance gains in recent research, establishing it as a key technical pathway for the advanced optimization of multimodal large models [25], [192], [158].

7.1 Task-supervised Fine-tuning

While specific implementations of task-supervised fine-tuning may incorporate technical variations, such as reinforcement learning or parameter-efficient strategies like LoRA (Low-Rank Adaptation) [320], the underlying principles are largely consistent with established fine-tuning practices for MLLMs. Accordingly, this paper examines the fine-tuning mechanisms of unified models from a task-oriented perspective. Based on the nature of the supervision signals, we categorize approaches by the task types present

in the fine-tuning data, leading to two primary paradigms: general-task fine-tuning and multi-task fine-tuning.

7.1.1 General-task Fine-tuning

General-task fine-tuning is a direct and widely adopted approach for developing unified models. This method employs a unified training objective to optimize the model end-to-end, aiming to comprehensively enhance its capabilities across a diverse range of tasks. Unlike single-task optimization, general-task fine-tuning utilizes a mixed dataset that integrates various task types, typically constructed by combining instruction-formatted data from multiple sources (e.g., image captioning, visual question answering, and text generation). This strategy ensures the training set covers a broad spectrum of multimodal tasks, thereby improving the model's instruction-following ability and cross-task generalization while maintaining a consistent processing pipeline.

(1) Data Composition

At the data level, the training set for general-task fine-tuning is typically formed by systematically combining data from multiple tasks. Based on differences in data sources and processing strategies, the construction methods can be categorized into two main technical pathways:

The first pathway involves fine-tuning on data derived from the refined processing of the pre-training dataset. This strategy combines differentiated hyperparameter configurations or training objectives, improving model performance by filtering and re-weighting the data used during pre-training. For instance, models like VL-GPT [195], Janus-Pro [26], and BAGEL [22] utilize instruction-formatted data in both pre-training and fine-tuning, with some overlap in the datasets. BAGEL, for example, adopts a progressive optimization strategy during fine-tuning, adjusting parameters such as image resolution and interleaved data ratios while constructing high-quality data subsets to achieve multi-stage performance enhancement. Similarly, methods like Chameleon [145] and Show-o [116] modify the weights of pre-training data during the fine-tuning stage, incorporating higher-quality multimodal and instruction data subsets without altering the fundamental training objective.

The second pathway involves augmenting the pre-training corpus with new, highly structured, and task-specific datasets. In this approach, the fine-tuning data often differs significantly in format and structure from the data used during pre-training. Such data is typically sourced from curated instruction-following datasets, such as LLaVA-1.5 [56] for multimodal dialogue or SEED-Data-Edit [318] for image editing, which provide explicit task instructions and annotated responses, as detailed in Sec. 6.2.2. For instance, Unified-IO2 [138] utilizes standard multimodal data pairs during pre-training but transitions to structured instruction data for fine-tuning, integrating tasks such as natural language processing, image and audio generation, and image understanding. Similarly, models like AnyGPT [140] and NExT-GPT [131] construct modally interleaved instruction datasets by amalgamating various tasks to enhance their any-to-any modal conversion capabilities. The meticulous curation of these multi-task datasets is instrumental in balancing and optimizing model performance across a diverse range of applications.

(2) Training Strategy

The training objectives for general-task fine-tuning are largely consistent with those employed during pre-training, with the autoregressive paradigm based on NTP remaining the dominant approach. However, specific implementations often incorporate distinct technical innovations. For instance, CM3 [212] introduces three differentiated fine-tuning schemes—standard, adversarial, and prompt-based—to optimize performance through fine-grained hyperparameter adjustments. Moreover, mirroring the pre-training phase, staged training strategies are also widely adopted in fine-tuning. Models such as BAGEL [22], Chameleon [145], and Show-o [116] utilize multi-stage fine-tuning processes, achieving progressive performance enhancements by systematically adjusting data configurations and training strategies at each stage.

Pros and Cons. General-task fine-tuning simplifies the training process through a unified optimization objective and synergistically enhances the model's comprehensive performance across multi-dimensional tasks, thereby achieving a balance between understanding and generation capabilities. However, this approach faces inherent challenges. First, intrinsic conflicts between the optimization objectives of different tasks often result in the performance of a generally fine-tuned model being inferior to that of a specialized model on specific tasks. Second, the addition of more task types can lead to catastrophic forgetting, making it a technical challenge to precisely balance performance across various tasks. How to optimize performance on specific tasks while ensuring the model's overall generality remains a core problem to be addressed in this field.

7.1.2 Multi-task Fine-tuning

In contrast to general-task fine-tuning, multi-task fine-tuning employs differentiated training strategies tailored to different tasks. This approach aims to mitigate potential inter-task conflicts that can arise from a unified training objective, thereby maximizing the model's performance on each individual task. Due to the inherent differences in the optimization objectives of various tasks, a unified training approach can make it difficult to achieve specialized expertise in specific domains.

(1) Task Types

Multi-task fine-tuning is a prevalent strategy for optimizing unified models for specific downstream applications. This approach is typically implemented via two primary pathways: (1) imbuing the model with new, domain-specific capabilities, such as medical image analysis, and (2) enhancing its performance on particular sub-tasks, such as personalized understanding or image editing. For instance, LLM-CXR [300] acquired specialized capabilities for processing chest X-ray images through instruction-based fine-tuning. Meanwhile, models like FROMAGe [299], OFA [123], Emu2 [137], and MM-Interleaved [139] have substantially improved their performance on specific multimodal sub-tasks by employing targeted fine-tuning objectives or parameter adjustments.

Furthermore, recent research has explored a paradigm that decouples understanding and generation tasks during the fine-tuning process, which represents an important

research direction in multi-task fine-tuning. For example, UniCTokens [171] implements a three-stage fine-tuning process on a frozen Show-o model backbone, with distinct optimization stages for warm-up, understanding, and generation. The warm-up stage stabilizes model parameters and adapts the backbone to the new training environment, typically using a low learning rate and limited data to ensure a smooth transition. Similarly, after an initial joint pre-training on mixed data, UniFork [307] independently optimizes for understanding and generation tasks during fine-tuning, freezing shared parameters to preserve the model's general capabilities.

(2) Task Differentiation

The differentiation of tasks within the multi-task fine-tuning paradigm is primarily realized through two key mechanisms: the strategic composition of training data and the design of task-specific training objectives.

At the data level, task differentiation is primarily achieved through the strategic composition of training corpora. As previously discussed, the independent fine-tuning of understanding and generation tasks exemplifies this approach by utilizing distinct datasets tailored to each function. Additionally, some studies adopt an incremental data strategy, introducing new, task-specific data to equip the model with additional capabilities. For instance, Janus-4o [178] is fine-tuned on the ShareGPT-4o-Image dataset, which includes both Text-to-Image and Text-and-Image-to-Image tasks. The inclusion of the latter successfully imparts novel image editing functionalities to the model.

At the training objective level, task differentiation is realized through the design of specialized loss functions. Even when training on mixed-data, some methods devise independent optimization objectives for different tasks. For example, CoDi-2 [136] is trained on a mixed dataset but applies separate loss functions for different outputs: a generation loss for images, a feature mean squared error for alignment, and a text prediction loss for text. Similarly, LLMBind [189] combines a text autoregressive loss, a segmentation mask loss, and a MoE auxiliary loss, while Minigpt-5 [134] jointly optimizes a text loss and a latent space diffusion loss on interleaved visual-text data. The use of composite loss functions is a core technical feature of multi-task fine-tuning, enabling the model to simultaneously optimize multiple performance dimensions by integrating diverse optimization targets.

Pros and Cons. Multi-task fine-tuning allows for the design of specialized optimization strategies tailored to the characteristics of different tasks, effectively mitigating inter-task conflicts and typically achieving superior performance on specific domain tasks. This method also enables more fine-grained control over capabilities through the design of differentiated loss functions. However, the implementation complexity of multi-task fine-tuning is significantly higher than that of general fine-tuning, requiring more engineering effort and domain-specific expertise. Furthermore, training costs may increase with the number of tasks, making it challenging to apply in resource-constrained scenarios.

With the maturation of unified task datasets, the technical barrier for general-task fine-tuning has been substantially lowered, making it the predominant strategy for building foundational unified models. Consequently, multi-task fine-tuning is increasingly reserved for adapting models to specialized downstream applications. This has led to a developmental paradigm where models are first endowed with broad, general-purpose capabilities and then refined for domain-specific expertise, thereby advancing the overall development of unified model fine-tuning.

7.2 Alignment Fine-tuning

In recent years, alignment fine-tuning has gained attention as an emerging training paradigm. This method further optimizes model performance by introducing human preference signals after the model has acquired foundational unified capabilities. Compared to traditional supervised fine-tuning, the core characteristic of alignment fine-tuning is its use of human preferences as the optimization objective, achieving deep alignment with human expectations through strategies such as DPO [321] and GRPO [322].

7.2.1 Implementation Details

In the alignment fine-tuning of existing unified models, two primary methods are employed: DPO and GRPO. The following section provides a detailed exposition of these methods' implementations and their variants.

DPO Strategy. This represents the most commonly adopted reinforcement learning approach in alignment fine-tuning for unified models, which is typically implemented by constructing training datasets D from human preferences. For instance, Emu3 [25] employs the DPO framework and utilizes manually annotated preference triplet data (prompt, preferred response, and rejected response) for optimization training. The training objective can be formalized as:

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_l|x)\cdot\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)\cdot\pi_\theta(y_w|x)}\right)\right], \quad (36)$$

where (x, y_w, y_l) represent the prompt, preferred response, and rejected response, respectively; π_θ and π_{ref} are the policies of the current and reference models; $\sigma(\cdot)$ is the sigmoid function; and β is a temperature parameter.

In addition, in order to better fit the training of the model, some variants have also evolved. For example, M2-omni [192] adopts a more sophisticated hybrid training strategy that combines DPO with instruction fine-tuning, leveraging LoRA for parameter-efficient updates. The training objective of this method combines the instruction-following loss and the preference alignment loss:

$$\mathcal{L}_{hybrid} = \lambda_1\mathcal{L}_{SFT} + \lambda_2\mathcal{L}_{DPO}, \quad (37)$$

where λ_1 and λ_2 are weighting coefficients, typically in the range $[0, 1]$, that balance the relative importance of the supervised fine-tuning loss ($\mathcal{L}_{SFT}$) and the human preference alignment loss ($\mathcal{L}_{DPO}$).

To further enhance alignment, some studies have adopted an iterative training paradigm. For instance, VARGPT-v1.1 [167] implements a progressive optimization strategy that alternates between supervised fine-tuning and DPO following an initial SFT phase. This method has demonstrated significant improvements, particularly

in complex generative tasks such as visual editing. This iterative process can be formalized as:

$$\theta_{t+1} = \text{DPO}(\text{SFT}(\theta_t, \mathcal{D}_{SFT}^{(t)}), \mathcal{D}_{pref}^{(t)}), \tag{38}$$

where t denotes the iteration round, and $\mathcal{D}_{SFT}^{(t)}$ and $\mathcal{D}_{pref}^{(t)}$ are the supervised fine-tuning and preference datasets for the t-th round, respectively. This iterative paradigm alternates between supervised fine-tuning and preference-based optimization, progressively refining the model's outputs to better align with human expectations.

GRPO Strategy. Some recent works have also employed the Group Relative Policy Optimization (GRPO) reinforcement learning strategy to train models, which abandons the critic model that is typically of the same size as the policy model. Formally, its optimization objective can be expressed as:

$$\mathcal{L}_{GRPO}(\theta) = \mathcal{L}_{clip}(\theta) - \beta \cdot D_{KL}(\pi_\theta \| \pi_{ref}), \tag{39}$$

where $\mathcal{L}_{clip}(\theta)$ represents the policy loss, and $D_{KL}(\pi_\theta \| \pi_{ref})$ denotes the KL divergence penalty.

The policy loss serves as the core driving force of GRPO, directly utilizing advantage values to guide the model's update direction. It can be formulated as:

$$\mathcal{L}_{clip}(\theta) = \mathbb{E}\left(\min\left(r_t(\theta) \cdot A_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) \cdot A_t\right)\right), \tag{40}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{ref}(a_t|s_t)$ is the probability ratio, reflecting the probability difference between the new policy and the old policy for selecting the same action in a specific state. A_t is the advantage function, representing how much better taking a particular action is compared to the average level in a specific state. The $\text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)$ function constrains the probability ratio within the interval $[1-\varepsilon, 1+\varepsilon]$ defined by the hyperparameter ε.

$D_{KL}(\pi_\theta \| \pi_{ref})$ is the Kullback-Leibler divergence between the current policy (π_θ) and the reference policy (π_{ref}), serving as a penalty term to prevent excessive policy update magnitudes. Specifically, it can be expressed as:

$$D_{KL}(\pi_\theta \| \pi_{ref}) = \mathbb{E}_{o_i \sim \pi_{ref}}\left[\log \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)}\right], \tag{41}$$

where q represents the query, and o_i refers to the i-th output in a set of sampled outputs.

GRPO provides new insights for reinforcement learning in training unified models, and has been commonly employed in recent works [323], [179], [181] for enhancing model generation performance during alignment fine-tuning. In recent works, GRPO-based alignment fine-tuning has been successfully applied to dual tasks in UniPic-2.0 [182] and equipped with a multi-dimensional reward system to ensure robust performance in Emu3.5 [210].

7.2.2 Terminological Clarification

In the literature, the term "post-training" often refers to an additional optimization phase following initial model training, which may include human preference alignment or further supervised fine-tuning [324], [325]. In contrast, "alignment fine-tuning" more narrowly denotes optimization guided by human feedback signals, such as DPO [321], which aligns model outputs with human expectations.

To avoid the definitional ambiguity and potential confusion associated with "post-training", this paper adopts the more specific term "alignment fine-tuning" to refer exclusively to model optimization based on human preference signals. This choice clarifies the technical scope and accurately reflects the core characteristics of these methods. This terminology will be used consistently throughout the remainder of this paper.

7.2.3 Pros and Cons

Human preference alignment methods offer a dual outcome. On one hand, alignment fine-tuning significantly enhances the consistency of model outputs with human expectations. The integration of human feedback enables more precise behavioral control, which is particularly effective for mitigating model-generated fabrications (i.e., "hallucinations") and improving the veracity of outputs in interactive tasks like dialogue generation. On the other hand, this approach presents notable challenges. A primary drawback is the substantial cost associated with collecting human preference data, which is both resource-intensive and time-consuming. Furthermore, reliance on human feedback introduces the risk of embedding annotator biases—such as cultural, regional, or demographic perspectives—into the model. This can inadvertently produce outputs that reflect the subjective preferences of the annotators, potentially compromising the model's objectivity and fairness [325], [326].

8 Training Data

The success of UFMs is fundamentally determined by the scale, quality, and diversity of their training data. The training data for these models is not a distinct category but rather a curated amalgamation of datasets traditionally used for both multimodal understanding and generation tasks. This section provides a comprehensive overview of the training data, detailing the data sources (Sec. 8.1), filtering methodologies (Sec. 8.2), construction techniques (Sec. 8.3), and a survey of existing datasets (Sec. 8.4).

8.1 Data Sources

The data used to train the UFM is drawn from four principal sources: large-scale internet crawls, curated public datasets, proprietary in-house collections, and synthetic generation. Each source provides a distinct profile regarding scale, quality, and diversity, and their combination is fundamental to the development of robust and capable models.

Internet Data. This represents the most extensive and readily available resource, sourced from large-scale web crawls of image-text pairs. These datasets provide unparalleled scale and diversity, capturing a vast spectrum of real-world scenarios and concepts. However, they are also subject to inherent challenges, including significant noise, inconsistent image-text alignment, and the potential presence of biased, unsafe, or low-quality content. The LAION-5B [327] dataset serves as a prime example, containing 5.85 billion image-text pairs filtered from the Common Crawl web data, and has become a foundational resource for pre-training prominent multimodal models such as Stable Diffusion. In contrast, Google's Conceptual Captions series (e.g., CC3M [314] and CC12M [328]) exhibits superior

alignment quality, as its alternative text is subjected to more sophisticated cleaning and simplification algorithms. Similarly, SBU Captions [329], comprising approximately one million images with descriptive titles from Flickr [330], was a foundational dataset frequently employed in early multimodal research.

Public Data. These datasets, often curated for specific visual or multimodal tasks, are characterized by their high-quality annotations and precise image-text alignment. These attributes make them particularly suitable for foundational model training and as "seed data" for instruction tuning. Despite their inherent limitations in scale, domain, and diversity, they provide significant value. For example, Microsoft COCO [312], which includes approximately 82,783 training images, each annotated with about five independent human-generated captions, is a widely recognized benchmark for image captioning and visual question-answering tasks. Similarly, datasets such as InstructPix2Pix [270], containing 450,000 image editing instructions, are instrumental in developing a model's image manipulation capabilities.

In-house Data. In-house data, often referred to as proprietary collections, constitutes another critical source, primarily utilized by large technology corporations. These datasets are distinguished by their immense scale and unique, domain-specific attributes. However, their proprietary nature renders them inaccessible to the broader research community. For instance, Meta leverages its extensive repository of public image-text posts from platforms such as Facebook and Instagram for model training [166]. Similarly, Google utilizes its indexed web data, YouTube video content, and, in compliance with privacy policies, Google Photos user data to train its internal models, including Gemini [11].

Synthetic Data. This emerging data source is created through the use of generative models, such as LLMs and text-to-image models, to synthesize novel image-text data. This approach allows for controlled, on-demand data generation, which can enhance diversity and address long-tail distribution challenges. For instance, LLaVA [55] utilizes GPT-4 to generate complex instructional data from COCO image labels, while SEED-Data-Edit [318] synthesizes image pairs for generation tasks. However, this method has notable limitations. The quality of synthetic data is constrained by the capabilities of the generative models, potentially leading to a "model-feeding-model" cycle. Furthermore, synthetic data may inherit and amplify biases from the source models and can lack real-world fidelity, necessitating careful evaluation before use.

8.2 Data Filtering

Data filtering is a critical process for refining raw data into high-quality training sets, a step that directly determines the model's ultimate performance. For UFMs, this process is particularly demanding as it needs to ensure both the accuracy of annotations for comprehension tasks and the alignment of instructions with outputs for generation tasks.

8.2.1 Filtering Methods

A high-quality training dataset for UFMs typically undergoes several filtering stages to ensure data accuracy, relevance, safety and others:

Heuristic Filtering based on Data Attributes. This fundamental step involves applying rule-based filters based on basic data properties. Common criteria include text length, image dimensions, and language identification. Additionally, deduplication of both images and text is performed to eliminate redundant data. For instance, the construction of the COYO-700M [331] included a rigorous deduplication process that filtered images by size, resolution, and aspect ratio. It also employed image hashing and feature vectors to remove duplicate or visually similar images, thereby improving training efficiency and model generalization.

Model-based Aesthetic and Quality Filtering. This approach utilizes pre-trained models to evaluate the aesthetic and technical quality of images. Aesthetic scoring models, such as the Aesthetics Predictor [332], are trained to predict human aesthetic preferences from image features. These models are then used to rate and filter out images that are blurry, poorly composed, or otherwise aesthetically deficient. A notable example is the LAION-Aesthetics-v2 [333] dataset, which was created by applying an aesthetic scoring model to the LAION-5B and retaining only the subset of images that surpassed a predefined aesthetic threshold. Images filtered in this manner, often characterized by superior composition and visual appeal, are highly valuable for fine-tuning text-to-image models. This filtering stage is crucial for curating high-quality datasets for generative tasks.

Modality Correlation Filtering. This method ensures strong alignment between different data modalities, most commonly by using a pre-trained vision-language model like CLIP [31] to compute a similarity score (e.g., CLIP-Score [334]) for image-text pairs. The score is calculated by encoding the image and text into feature vectors and then computing their cosine similarity. A predefined threshold is used to discard pairs with weak semantic alignment. For instance, the LAION-5B [327] dataset was constructed by applying CLIPScore filtering (with a threshold between 0.28 and 0.32) to the Common Crawl. This technique is also applied to other modalities; the BAGEL dataset [22], for example, uses visual similarity to filter video clips, while SEED-Data-Edit [318] employs correlation filtering to align image generation data with instructions. This filtering step is fundamental to improving the quality of modal alignment in training data.

Content Safety and Compliance Filtering. This critical filtering stage employs classifiers and keyword matching to identify and remove unsafe content, such as Not-Safe-For-Work (NSFW) material, which includes depictions of violence, pornography, and hate speech. Beyond safety, this process also addresses compliance issues related to data privacy, copyright, and legal restrictions. It aims to ensure that the dataset is free from personally identifiable information (PII), unauthorized copyrighted material, and other content that may violate regulatory or ethical standards. For instance, the LAION team utilized an NSFW detection model [335] to apply safety labels to its data, enabling users to filter out potentially inappropriate content. This type of screening is essential for enhancing both the safety and legal compliance of the training data.

8.2.2 Filtering Pipeline

The following section details a comprehensive, multi-stage pipeline that illustrates a complete, end-to-end process for data filtering. This process is designed to refine high-quality training samples from massive raw datasets and typically proceeds through the following core stages:

(1) Preliminary Data Cleaning

This initial stage is dedicated to eliminating invalid and low-utility data through fundamental validation checks and attribute-based filtering.

- **Deduplication and Format Validation:** The process begins with the removal of redundant data, identified via duplicate content hashes (e.g., MD5, SHA-1) or source URLs. Subsequently, format validation is performed using standard libraries (e.g., Python's Pillow, ImageMagick) to verify image file integrity and confirm adherence to standard parsable formats (e.g., jpeg, png), leading to the exclusion of corrupted or unreadable files.
- **Attribute-based Filtering:** A preliminary screening based on metadata is conducted to discard images with low informational value. This involves filtering out images that do not meet a minimum resolution threshold (e.g., 256×256 pixels) or that exhibit extreme aspect ratios (e.g., >3:1) [336], [337] . Concurrently, associated text descriptions are filtered by length to remove entries that are excessively short (e.g., fewer than three words) or overly long (e.g., more than 100 words), as such texts can introduce noise or cause semantic drift.

Upon completion of this stage, the dataset is purged of corrupted files, duplicates, and entries with unsuitable physical attributes, thereby establishing a baseline of data integrity.

(2) Content Quality and Safety Filtering

This stage utilizes automated models to filter content based on safety, compliance, and quality criteria, removing materials that are unsafe, non-compliant, or of low aesthetic and technical value.

- **Safety and Compliance Screening:** Dedicated classifiers, such as NSFW models [335], are applied to identify and remove inappropriate content, including pornography and violence. Concurrently, text classifiers or keyword-based filters are used to screen for hate speech and other non-compliant material. For instance, the LAION project [327] implements this process by appending safety labels to its data, enabling downstream filtering.
- **Aesthetic and Technical Quality Filtering:** Pre-trained aesthetic scoring models [332] are employed to evaluate and discard images with low aesthetic value. In parallel, technical quality assessment models filter out images exhibiting defects such as blurriness, poor lighting, excessive noise, or prominent watermarks.

Upon completion of this stage, the dataset is refined to exclude unsafe and low-quality samples, thereby ensuring regulatory compliance and elevating its technical standard.

(3) Modality Alignment Filtering

This pivotal stage employs pre-trained vision-language models to enforce strong semantic correspondence between images and their associated textual descriptions.

- **Image-Text Relevance Scoring:** A foundational step in this stage involves quantifying the relevance between modalities. Pre-trained vision-language models, such as CLIP, are employed to compute a similarity score (e.g., CLIPScore [334]) for each image-text pair.
- **Threshold-based Filtering:** An empirically determined threshold is applied to the relevance scores. For instance, the LAION-5B [327] utilized a CLIPScore threshold between 0.28 and 0.32. Pairs with scores below this threshold are discarded. This process is instrumental in eliminating noisy data characterized by weak or irrelevant image-text alignment, thereby enhancing overall dataset quality.

Upon completion of this stage, the dataset is refined to contain only highly relevant and well-aligned multimodal pairs, significantly improving its semantic coherence.

(4) Advanced Deduplication and Data Balancing

This final stage further refines the dataset by eliminating semantically redundant content and, where necessary, rebalancing the data distribution to mitigate potential biases.

- **Semantic and Visual Deduplication:** This step employs advanced deduplication techniques to identify and remove near-duplicate entries. For visual data, methods such as deep feature-based similarity metrics are utilized to detect and eliminate visually similar images. For textual data, deduplication is performed based on n-gram overlap or the cosine similarity of sentence embeddings. For instance, the construction of the COYO-700M dataset [331] incorporated a rigorous visual deduplication phase to enhance data diversity.
- **Data Balancing (Optional):** In certain application contexts, data balancing may be performed to prevent the model from developing biases toward high-frequency concepts [336]. This is typically achieved by down-sampling over-represented categories or up-sampling under-represented ones.

Upon completion of this stage, the dataset is characterized by its diversity, balance, and absence of redundant samples, rendering it suitable for high-quality model training.

The data filtering pipeline detailed above refines large-scale raw web data into a training set characterized by appropriate scale, content safety, high quality, and strong image-text relevance. This process establishes a robust data foundation for the development of high-performance unified models. It should be noted, however, that many datasets are constructed by leveraging pre-existing curated data, which may obviate the need to repeat certain filtering stages.

8.3 Data Construction

Data construction is the process of transforming raw multimodal data into a format suitable for training UFMs, a step that is critical to their successful development. Unlike traditional single-task datasets, UFMs are required to possess both comprehension and generation capabilities. Consequently, their training data should encompass a diverse array of modal interaction patterns and task formats. The primary challenge in data construction is to uniformly integrate raw data from disparate sources and with varying structural characteristics, enabling it to yield maximum learning efficacy within a unified training framework. This process should not only preserve the semantic integrity of the original data but also ensure that the constructed data can effectively drive performance improvements across

various downstream tasks. Specifically, data construction for unified models primarily involves the following methods.

8.3.1 Conversion from Public Datasets.

This method involves the systematic transformation of public academic datasets into a standardized format suitable for unified model training. The process reformats data from various single-task datasets (such as VQA [338] for visual question answering, and RefCOCO [339] for referring expression comprehension) into a uniform `<instruction, input, output>` structure. This standardization enables the integration of heterogeneous data sources into a single, cohesive training framework.

For example, an original annotation from the COCO dataset, such as "a man throwing a frisbee", can be reformatted into an instructional format for a comprehension task:

```
{"instruction": "Describe the content of the
    image.",
 "input": <image>,
 "output": "This image shows a man throwing a
    frisbee on the grass in a park."}
```

Conversely, the generation data can be structured as:

```
{"instruction": "Generate an image based on
    the following description.",
 "input": "A man throwing a frisbee on the
    grass in a park.",
 "output": <image>}
```

This bidirectional conversion allows a single data source to support the training of both comprehension and generation.

Beyond standard instruction formats, this conversion methodology is also applied to construct interleaved data. For instance, the BAGEL [22] reconstructs filtered video clips into `<video, caption, video, caption>` sequences to enhance the model's ability to process temporal multimodal information. Similarly, web data can be converted into `<text, caption, image>` interleaved formats, which more accurately simulate the natural distribution of multimodal information in real-world contexts. Interleaved formats are particularly beneficial for tasks requiring contextual understanding across multiple modalities, such as document analysis, video narration, or interactive storytelling, as they enable models to learn the relationships and dependencies between sequential multimodal elements.

While this conversion approach is straightforward and effective for rapid model prototyping, it is constrained by the scope and quality of the original datasets. This dependency can limit the diversity and novelty of the training data. A common strategy to mitigate this is to use human-annotated data as a high-quality seed set for pre-training or validation, while leveraging automatically generated or web-crawled data to expand the dataset's scale. For example, the LLaVA-Instruct dataset [55] used human-annotated COCO captions as a foundation and then expanded its instruction data using GPT-4, thereby achieving a balance between quality and scale.

8.3.2 Generation using Large Models.

The use of powerful large models, such as GPT-4o [10] and LLaVA [55], for synthetic data generation has become a prominent and highly efficient method for data construction. This approach leverages advanced LLMs and multimodal models that can process and generate multimodal data. The core principle involves providing a model with foundational visual information, such as image content and associated labels, and then prompting it to generate diverse and complex data formats, including dialogues, question-answer pairs, and detailed instructions. This methodology enables the cost-effective, large-scale generation of high-quality and varied instruction-following data.

For example, given an image of a cat with the labels *"cat, sofa"*, an MLLM can be prompted to generate various structured data formats. These can range from simple conversational data (e.g., *{"user": "What is this cat doing?", "ai": "It is resting on the sofa."}*) to complex reasoning data (e.g., *{"user": "If I wanted to adopt this cat, what should I consider?", "ai": "Based on its relaxed posture, it likely requires a quiet and comfortable living environment..."}*). The LLaVA-Instruct-150K dataset [55] is a notable example of this technique; it was created by using GPT-4 to generate 150,000 high-quality multi-turn conversations, detailed descriptions, and complex reasoning questions from COCO images, establishing a foundational resource for subsequent research.

This generation-based approach also extends to the creation of datasets for generative tasks. For instance, ShareGPT-4o-Image [178], a large-scale image generation dataset, contains 91,000 prompt-image pairs where all images were synthesized by GPT-4o [10]. Furthermore, large models are employed to augment existing datasets. A common strategy involves using a model like GPT-4o to generate data that addresses known weaknesses or underrepresented concepts, thereby enhancing the scale, diversity, and robustness of the training data.

While this method offers significant advantages in efficiency and scalability, its limitations should be acknowledged. The quality of the synthetic data is inherently dependent on the capabilities of the underlying generative model. Consequently, the generated data may inherit biases, factual inaccuracies, or other limitations from the source model. This creates a risk of a "model-feeding-model" cycle, where errors and biases are amplified, potentially compromising the authenticity and reliability of the final training dataset.

8.3.3 Human Annotation & Crowdsourcing.

Although human annotation and crowdsourcing methods exhibit significant limitations in terms of cost investment, time efficiency, and data scale, they possess irreplaceable advantages in data quality and human cognitive consistency. This approach can produce high-quality annotated data that most closely aligns with human thinking patterns and expression habits, providing crucial quality benchmarks for model training. The COCO Captions dataset [312] serves as a quintessential example of human annotation, where each image in COCO is equipped with five independent human-written descriptions. Its exceptional annotation quality has established it as a classical data source for multimodal model training. Additionally, many datasets employ semi-supervised human annotation strategies. For instance, SEED-Data-Edit [318] implemented human re-annotation processing for real-world scenario data during its construction. Although the majority of its data relies on automatic

model generation, critical quality control steps still depend on human intervention.

In recent years, the rise of human preference alignment methodologies has catalyzed the development of novel human-annotated datasets. Human preference data is primarily collected through user behavior during interaction processes, combined with acceptance or rejection mechanisms in model evaluation to achieve precise alignment with human expectations. This evaluation framework typically relies on specialized human preference metrics, such as ImageReward [340] and HPSv2 [341], which can effectively quantify human preference levels for model-generated content, thereby guiding models to produce outputs that better conform to human expectations.

Despite the exceptional performance of human annotation and crowdsourcing methods in data quality and human expectation consistency, their inherent efficiency bottlenecks and scale limitations make them inadequate for meeting the urgent demand for massive data in large-scale unified model training. Therefore, in practical applications, researchers typically utilize human-annotated data as high-quality seed data or quality benchmarks, combining them with other data acquisition methods to achieve optimal balance between quality and scale.

8.4 Existing Datasets

The training data for unified models encompasses a diverse array of formats, extending well beyond simple image-text pairs. The complexity and structural diversity of these datasets are foundational to the models' advanced capabilities. As illustrated in Fig. 15, the training data can be broadly categorized into several principal types, which are detailed in the following subsections.

Image Datasets. Image datasets form the foundational modules of multimodal learning, primarily consisting of images paired with annotations. These range from basic RGB images to complex data formats such as depth maps and optical flow, each with corresponding annotations. For instance, the NYU Depth Dataset V2 [466] provides 1,449 densely annotated RGB-depth image pairs, offering crucial data for 3D scene understanding and depth estimation. Large-scale classification datasets have shown significant growth, evolving from ImageNet 2012's [467] 1.3 million images to ImageNet-21K's [311] 14.19 million images across 21,841 categories, reflecting the escalating data requirements of computer vision tasks. Scene understanding datasets, such as the Places365 [468] series, have also scaled, expanding from a standard version with 1.8 million images to a challenge version with 8 million, focusing on the precise recognition of 365 distinct scene categories. Furthermore, specialized datasets for optical flow estimation, like Flying Chairs [471] (22,232 images with corresponding flow annotations) and MPI Sintel [474] (23 training sequences with ground-truth optical flow), provide essential data for motion estimation and dynamic scene understanding in video sequences.

Image-Text Pair Datasets. Image-text pairs are a core data modality in multimodal learning, exhibiting remarkable scalability from millions to billions of pairs. The LAION series represents the largest collection of such pairs, scaling from LAION Aesthetics v2.5 [333] (176,000 pairs) to LAION-400M [433] (400 million pairs) and culminating in LAION-2B-en [327] (2.32 billion English pairs), providing vast resources for large-scale multimodal pre-training. In terms of fine-grained annotation, the MSCOCO [312] dataset offers 413,900 high-quality annotations for 82,800 training images, balancing quality with a substantial scale. Domain-specific datasets like Objects365 [441], with 2 million images and 30 million bounding box annotations, cater to the needs of object detection. Additionally, referring expression datasets such as the RefCOCO [339] series (RefCOCO, RefCOCO+, and RefCOCOg) provide crucial support for fine-grained vision-language understanding tasks. Recently, high-quality synthetic datasets like ShareGPT4V [463] (100,000 samples, extended to 1.246 million in ShareGPT4V-PT) have gained prominence, demonstrating the potential of synthetically generated training data.

Visual Question Answering (VQA) Datasets. VQA [338] datasets have evolved from simple question-answering to complex, multi-step reasoning, providing essential benchmarks for evaluating multimodal understanding and inference. Foundational datasets like VQA v2.0 [338] (443,800 questions) established the standard evaluation framework. More complex, reasoning-oriented datasets have significantly increased in scale and difficulty; GQA [417], with 22.67 million questions over 113,000 images, emphasizes scene graph reasoning and multi-step logic. Domain-specific extensions include ScienceQA [424] (12,700 multimodal multiple-choice questions), A-OKVQA [420] (17,100 question-answer pairs requiring external knowledge), and VizWizVQA [418] (31,000 real-world questions from visually impaired users). The video domain has also seen rapid development, with datasets like MSVD-QA [426] (30,900 QA pairs on 1,200 videos) and MSRVTT-QA [426] (158,600 QA pairs on 6,513 videos) providing benchmarks for video understanding and temporal reasoning.

HTML/XML Structured Datasets. Structured document datasets provide rich contextual information and realistic web environments for multimodal understanding. These are primarily sourced from large-scale web crawls and knowledge bases. Common Crawl News [415] is a representative example, containing 45 million documents (4,600 GB), 18 million unique images, and 121 billion text tokens, offering a vital resource for learning from real-world web-based multimodal information. The English Wikipedia dataset [416], a high-quality knowledge base, provides 16 million documents (3,830 GB), 7 million unique images, and 102 billion tokens, with its multilingual nature making it a key resource for cross-lingual multimodal learning. These datasets preserve the original layout and structural information of web pages, enabling models to learn the relationships between images, text, and their spatial arrangement.

Video Datasets. Video datasets feature diverse characteristics, ranging from short clips to long sequences and from simple actions to complex scenes, providing rich resources for video analysis. The large-scale YT-Temporal-180M [406], with 6 million videos and 180 million extracted frames, is one of the largest video datasets available. Action recognition datasets are central to video understanding; examples include UCF101 [410] (13,300 clips, 101 action classes) and

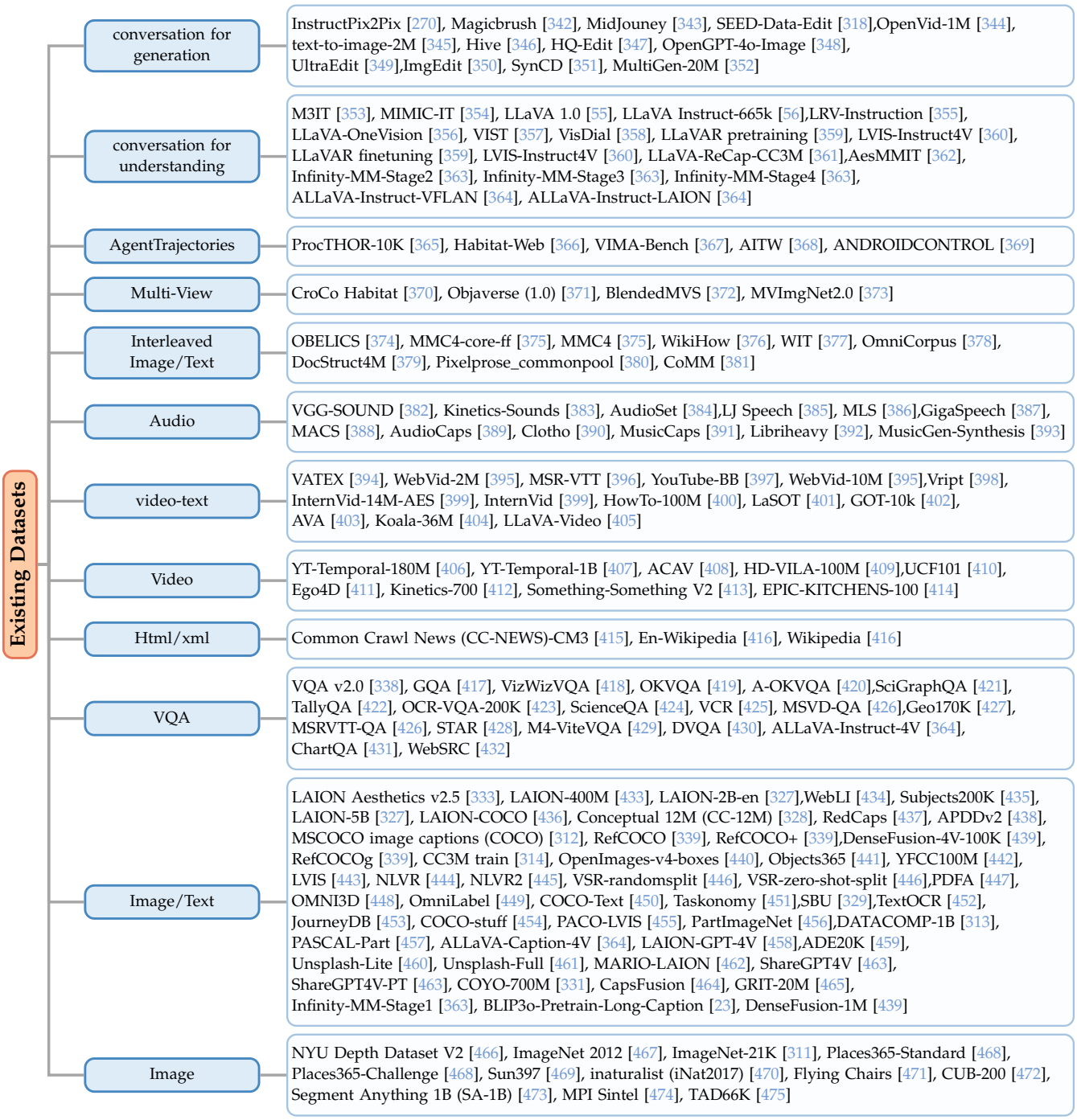

Fig. 15: A summary of existing datasets, categorized by data type, along with representative examples for each category. The existing datasets can be mainly divided into 12 data forms. In Sec. 8.4, different forms of data will be briefly described, and the appendix will provide a detailed statistical features to the datasets.

Something-Something V2 [413] (168,900 clips, 174 object interaction classes). Datasets for long-term and complex scenes, such as Ego4D [411] (3,670 hours of first-person video with multimodal annotations) and EPIC-KITCHENS-100 [414] (100 hours of kitchen activity video), present challenges for deep video understanding. Domain-specific datasets like Kinetics-700 [412] (545,300 clips, 700 action classes) are for large-scale action recognition.

Video-Text Pair Datasets. Video-text pair datasets are fundamental for tasks like video captioning, description generation, and cross-modal retrieval. Multilingual datasets like VATEX [394] (41,250 videos with 825,000 bilingual Chinese-English annotations) facilitate cross-lingual video understanding. Large-scale web-crawled datasets include WebVid-10M [395] (10.7 million video-text pairs, 52,000 hours) and WebVid-2M [395] (2.5 million pairs, 13,000 hours), providing abundant material for pre-training. Instructional video datasets like HowTo-100M [400] (136 million clips, 134,500 hours) are particularly valuable. Classic video description datasets such as MSR-VTT [396] provide

detailed English sentences for 6,513 video clips (20 descriptions per video). More recent datasets like InternVid-14M-AES [399] focus on aesthetic quality, collecting 14.5 million video clips with high aesthetic scores.

Audio Datasets. Audio datasets encompass various forms, from pure audio signals to audio-text pairs, supporting tasks in audio understanding, classification, and generation. Pure audio classification datasets like VGG-SOUND [382] (177,800 labeled clips) and Kinetics-Sounds [383] (15,000 ten-second clips) form a basis for audio event detection. Audio-text pair datasets are crucial for multimodal audio understanding; AudioSet [384] is highly influential, with 2.08 million human-verified 10-second clips annotated across 632 audio classes. Audio captioning datasets like AudioCaps [389] (46,000 clips with human-written descriptions) and Clotho [390] (2,893 samples with 14,465 annotations) provide high-quality data for audio captioning and retrieval. Music-specific datasets like MusicCaps [391] (5,521 audio-text pairs) are optimized for music description. Speech datasets such as LJ Speech [385] (13,100 short clips) and the Multilingual Librispeech (MLS) [386] dataset (44,500 hours of English and 6,000 hours in other languages) are vital for speech recognition and synthesis.

Interleaved Image-Text Datasets. Interleaved image-text datasets simulate the natural co-occurrence of images and text in real-world documents, providing essential resources for training models to understand complex multimodal content. These datasets preserve the spatial relationships and contextual continuity of the original documents. OBELICS [374] is a large-scale example, containing 141 million English documents with 115 billion text tokens and 353 million images sourced from web crawls. MMC4 [375] (Multimodal C4) is another key dataset, with 101.2 million documents featuring 571 million images interleaved within 43 billion English tokens. Interleaved versions of existing datasets like CC12M [328], CC3M [314], and RedCaps [437] have also been created. WikiHow [376], with 230,000 tutorial documents, organizes images and text in a natural instructional sequence, providing unique data for instruction-following tasks.

Multi-View Datasets. Multi-view datasets provide the data foundation for 3D scene understanding, stereo vision, and novel view synthesis. They typically contain images of the same scene or object from different viewpoints, along with corresponding geometric information. CroCo Habitat [370] uses a simulator to generate 1.82 million image pairs of the same scene from adjacent viewpoints, facilitating self-supervised stereo matching. Objaverse 1.0 [371] is a large-scale 3D object dataset with 818,000 models and 2.35 million labels, supporting 3D object recognition and generation. BlendedMVS [372] focuses on multi-view stereo tasks, providing 17,000 training samples with high-quality depth annotations and camera parameters for developing precise 3D reconstruction algorithms.

Agent Trajectory Datasets. Agent trajectory datasets record the behavioral sequences and decision-making processes of agents in various environments, providing crucial data for embodied AI and reinforcement learning. ProcTHOR-10K [365] offers 10,000 trajectories of agents navigating and interacting in procedurally generated indoor environments. Habitat-Web[366] is a larger-scale dataset with 293 million action records from 226,000 hours of real-world teleoperation. VIMA-Bench[367] provides 650,000 trajectories and multimodal demonstrations for learning complex manipulation tasks. These datasets capture not only action sequences but also environmental states, task goals, and outcomes, offering valuable learning resources for developing intelligent agents capable of performing real-world tasks.

Conversational Understanding Datasets. These datasets focus on multi-turn, interactive vision-language understanding tasks, providing a foundation for developing conversational multimodal systems. M3IT [353] (Multitask Multimodal Instruction Tuning) contains 2.4 million instances covering various instruction-formatted multimodal tasks. MIMIC-IT [354], a large-scale multimodal instruction tuning dataset, provides 2.8 million instruction-response pairs with images and videos. The LLaVA series has scaled from 158,000 instructions (LLaVA 1.0 [55]) to 665,000 conversational data points (LLaVA Instruct-665k [56]), driving rapid progress. Visual storytelling datasets like VIST [357] (16,000 stories with corresponding image sequences) and visual dialog datasets like VisDial [358] (123,000 multi-turn conversations about images) emphasize contextual understanding and interactive capabilities.

Image Generation Datasets. Image generation datasets are specifically designed for image editing, modification, and creation tasks. InstructPix2Pix [270] is a representative image editing instruction dataset with 450,000 examples, each containing an original image, an editing description, and a target image. Magicbrush [342] focuses on finer-grained editing, with 88,000 instructions for local edits and detail modification. Due to the scarcity of real-world data, synthetic instruction generation has become a primary data source. Datasets synthesized from image-text pairs often combine multiple sources, such as Unsplash [461], LAION-Aesthetics [333], JourneyDB [453], and OpenGPT-4o-Image [348], providing diverse training material for image generation and editing tasks.

In summary, the training data for UFMs encompasses a diverse array of data types that extend well beyond basic image-text pairs. This includes foundational unimodal datasets for images, videos, and audio, as well as complex multimodal formats such as visual question answering, interleaved documents, multi-view imagery, and agent trajectories. Such diversity is essential for developing comprehensive capabilities across understanding, generation, and reasoning tasks. A detailed summary of representative datasets, including their scale, sources, and language coverage, is provided in the appendix for reference.

8.5 Summary

The construction of training data for large multimodal models represents a complex and sophisticated systems engineering endeavor. This process begins with sourcing raw data from massive repositories such as LAION [327], followed by meticulous filtering based on CLIP Scores [334] and aesthetic models. The resulting data spans a wide spectrum of types, from basic image-text pairs to complex interleaved sequences like those found in MMC4 [375]. A pivotal stage is the construction of high-quality instruction data through methods like LLaVA-Instruct, which is essential for eliciting the model's advanced capabilities. The

process culminates in a comprehensive quality assessment using benchmarks like MME [476] and MMBench [477], thereby closing the data engineering loop.

Ultimately, innovation in data engineering serves as the core engine driving the evolution of multimodal model capabilities. Continuous advancements in the breadth of data sources, the precision of filtering methods, the richness of data types, the novelty of instruction construction, and the rigor of quality assessment are the direct determinants of the final performance of unified multimodal models.

9 Benchmark

UFMs have greatly advanced multimodal intelligence, necessitating comprehensive evaluation protocols. This section presents an organized introduction of existing benchmarks, structured into three main areas. First, Sec. 9.1 reviews benchmarks for model understanding, encompassing image (Sec. 9.1.1), video (Sec. 9.1.2), audio (Sec. 9.1.3), and mixed-modal (Sec. 9.1.4) understanding tasks. Next, Sec. 9.2 examines generative benchmarks, including those for image (Sec. 9.2.1), video (Sec. 9.2.2), audio (Sec. 9.2.3), and mixed-modal (Sec. 9.2.4) generation. Finally, Sec. 9.3 highlights recent progress in mix-modality generation, reflecting the integration of understanding and generation across complex multimodal scenarios. Fig. 16 categorizes existing benchmarks by modality and task, and selects representative works. Tab. 5 selects the most common benchmarks by modality and lists their data sources, evaluation indicators, and sample size.

9.1 Benchmark for Understanding

Understanding is a fundamental capability of MLLMs. Although foundational tasks such as question answering are common across modalities, evaluation criteria vary by input type. For example, evaluating video understanding requires assessing a model's ability to capture continuity and temporal dynamics across frames, which is distinct from static image inputs [567]. Similarly, evaluating the audio modality focuses on the model's understanding of non-visual information, such as speech, sounds, and music [609], [611], [612], [613]. This section overviews benchmarks for evaluating MLLMs' understanding across different modalities. We further categorize these benchmarks by the specific capabilities they target, such as Optical Character Recognition (OCR) [517], [518], [516], safety [562], and fine-grained perceptual or reasoning skills [537], [534], [535].

9.1.1 Image

Foundation Capability. Evaluating foundational perception and reasoning requires comprehensive benchmarks, which have evolved in both scope and methodology. Early benchmarks addressed specific challenges. For example, VizWiz [418] measures robustness in real-world scenarios by testing VQA models with low-quality images and diverse questions. VQA v2.0 [338] reduces language prior bias using a balanced dataset of complementary image pairs, improving the measurement of visual grounding.

Subsequent evaluations focus on comprehensive and objective frameworks. For instance, MME [476] proposes a multi-task framework using "yes/no" questions to assess perception and cognition while minimizing data contamination. MMBench [477] introduces a multi-dimensional, multiple-choice benchmark for assessing fine-grained capabilities and applies LLM-assisted answer-matching to enhance robustness. Recent benchmarks address previous data scale limitations. The SEED-Bench series [481], [482] provides large-scale, human-annotated multiple-choice datasets, with SEED-Bench-v2 adding hierarchical evaluation and image generation assessment. LVLM-eHub [478] integrates existing datasets for quantitative analysis and collects user preferences through an online arena platform. MME-RealWorld [494] offers a large-scale, human-annotated dataset of high-resolution images, using challenging real-world scenarios to reveal model limitations in practical applications. Recent research also investigates model failure mechanisms. MMVP [492] diagnoses failures in visual pattern comprehension via "CLIP-blind pairs," enabling analysis at the representation layer. In summary, the evaluation of image understanding is evolving toward more comprehensive, large-scale, and in-depth analyses.

World Knowledge. A key capability of MLLMs is leveraging world knowledge to interpret visual content. Benchmarks in this category evaluate a model's performance on visual question answering that requires implicit commonsense knowledge or explicit disciplinary knowledge from fields like science and the arts.

Several benchmarks test expert-level knowledge in academic domains. ScienceQA [424] evaluates MLLMs focus on scientific problems, such as biological or environmental science. MMMU [498] evaluate expert-level cognition by posing university-level, multidisciplinary problems. MMMU-Pro [501] enhances evaluation rigor by filtering questions solvable by text alone, compelling models to rely on visual evidence. To address linguistic and cultural diversity, other benchmarks have been developed. CMMU [499] and CMMMU [500] provide dedicated Chinese-language benchmarks, evaluating knowledge from primary school to university levels. Another direction aligns evaluation with practical, real-world applications. MDI-Benchmark [502] adopts a user-centric approach by covering scenarios from daily life and introduces an age-stratification dimension to assess personalized performance.

Reasoning and Math. Reasoning evaluation is a critical dimension for MLLMs. Unlike simple perception tasks, these benchmarks require multi-step logical reasoning over complex visual information, such as geometric figures.

Some benchmarks focus on specific paradigms of visual reasoning. For example, InfiMM-Eval [506] assesses open-ended, complex visual reasoning by categorizing problems into deductive, abductive, and analogical paradigms, and provides granular quality metrics by evaluating intermediate steps. Visual CoT [495] promotes visually grounded reasoning by providing a large-scale dataset with bounding box annotations for key evidence, enhancing model interpretability. M^{3}CoT [509] delivers a more rigorous benchmark for Multimodal Chain-of-Thought by introducing multi-domain, multi-step problems based on the ScienceQA [424].

In the specialized domain of mathematical reasoning, comprehensive evaluation suites have emerged. MathVista [534] offers a holistic benchmark by integrating 28

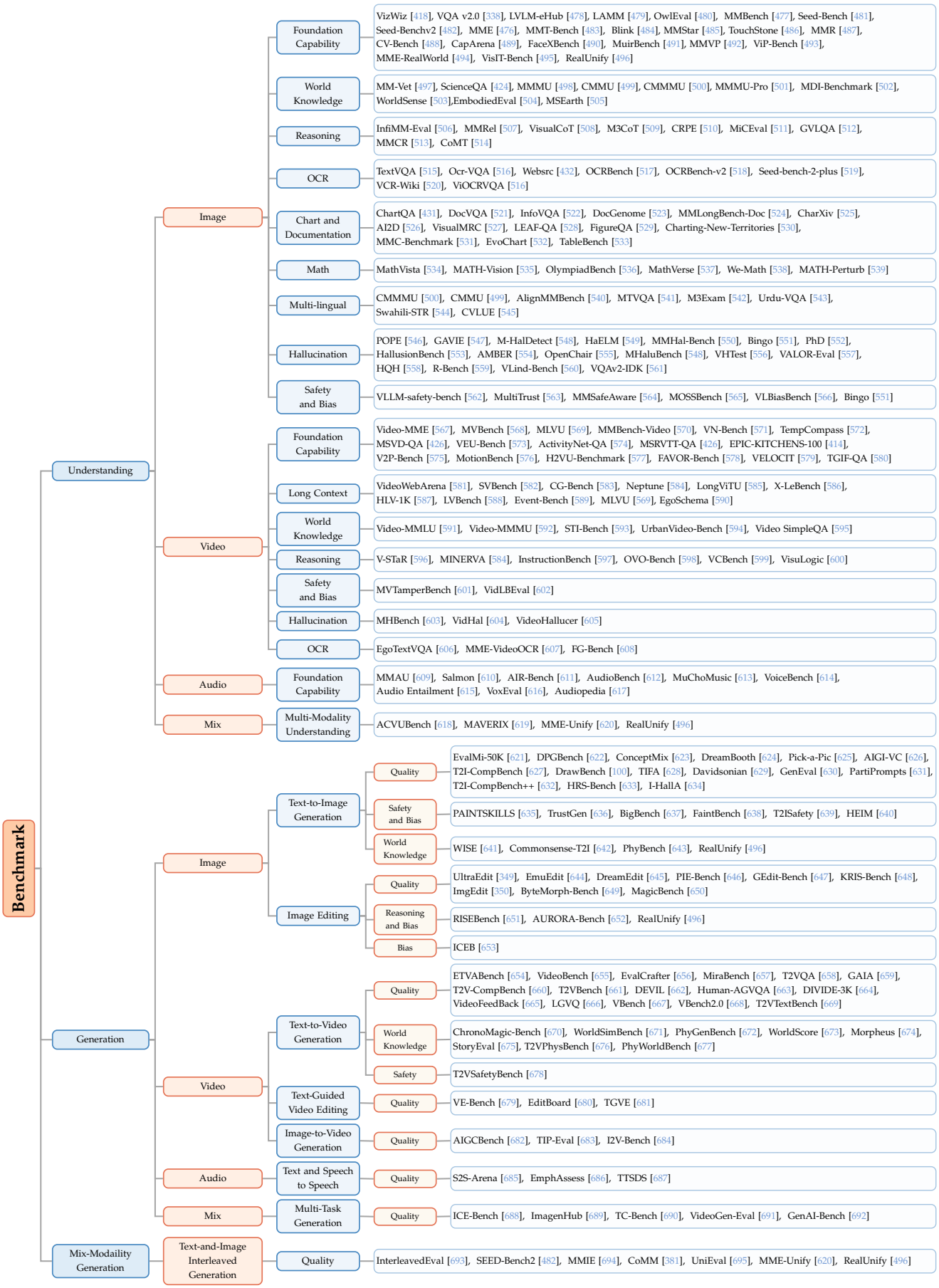

Fig. 16: Summary of multimodal benchmarks by modality and task, with representative benchmarks. The diagram highlights coverage across Image, Video, Audio, and Mix modalities, detailed in Sec. 9.

TABLE 5: Statistics of representative UFM benchmarks (grouped by modality). The table organizes Image/Video/Mix benchmarks and reports, for each, its task type, evaluation metrics, data sources, and scale.

Modality	Benchmark	Task	Metric	Data Source	QA Pairs
Image	EvalMi-50K [621]	T2I Generation	Human Evaluation	EvalMi-50K [621]	50k
	GenEval [630]	T2I Generation	Accuracy	GenEval [630]	2.2k
	WISE [641]	T2I Generation	MLLM Evaluation	WISE [641]	1k
	DPG-Bench [622]	T2I Generation	MLLM Evaluation	COCO [312], PartiPrompts [631], DSG-1K [629], Object365 [441]	1k
	ConceptMix [623]	T2I Generation	MLLM Evaluation	ConceptMix [623]	2.1k
	DALL-Eval [696]	T2I Generation	Accuracy, Human Evaluation	DALL-Eval [696]	7.1k
	Commonsense-T2I [642]	T2I Generation	MLLM Evaluation	Commonsense-T2I [642]	0.15k
	DreamBooth [624]	T2I Generation	DINO, CLIP-I, CLIP-T	DreamBooth [624]	3.0k
	Pick-a-Pic [625]	T2I Generation	PickScore	Pick-a-Pic [625]	0.5k
	DrawBench [100]	T2I Generation	Human Evaluation	DrawBench [100]	0.2k
	TIFA [628]	T2I Generation	Accuracy	COCO [312], DrawBench [100]	4k
	PhyBench [643]	T2I Generation	MLLM Evaluation, Human Evaluation	PhyBench [643]	0.7k
	EmuEdit [644]	Image Editing	CLIP-I, CLIP-T, L1-Distance, DINO, Human Evaluation	MagicBrush [650]	5.6k
	RISEBench [651]	Image Editing	MLLM Evaluation, Human Evaluation	RISEBench [651]	0.064k
	DreamEditBench [645]	Image Editing	Human Evaluation	DreamEditBench [645]	0.44k
	PIE-Bench [646]	Image Editing	PSNR, LPIPS, MSE, SSIM, CLIPSIM	PIE-Bench [646]	0.7k
	PIE-Bench++ [646]	Image Editing	MLLM Evaluation, AspAcc-CLIP	PIE-Bench [646]	0.7k
	ICEB [653]	Image Editing	SF, S2D, AL2D, HD	ImageNet [467]	4.2k
	MagicBrush [650]	Image Editing	L1-Distance, L2-Distance, CLIP-I, CLIP-T, DINO	COCO [312]	1.73k
	GEdit-Bench [647]	Image Editing	MLLM Evaluation	GEdit-Bench [647]	1.2k
	ImagenHub [689]	T2I Generation; Image Editing	LPIPS, CLIP-I, CLIP-T, Human Evaluation	MagicBrush [650], DreamEditBench [645]	1.1k
Video	ETVABench [654]	T2V Generation	MLLM Evaluation	VBench [667], EvalCrafter [656], PhyGenBench [672], ChronoMagic-Bench [670]	2k
	Video-Bench [655]	T2V Generation	MLLM Evaluation	Video-Bench [655]	0.42k
	DIVIDE-3K [664]	T2V Generation	MOS, MOSA, MOST	YFCC-100M [442], Kinetics-400 [412]	3.6k
	VideoFeedBack [665]	T2V Generation	VideoScore	VidProM [697]	0.76k
	LGVQ [666]	T2V Generation	UGVQ-Score	LGVQ [666]	0.46k
	VE-Bench [679]	Text-Guieded Video Editing	VE-Bench QA	DAVIS [698], Kinetics-700 [412], Sintel [699], Spring [699], Sora, Kling	0.17k
	BalanceCC [700]	Text-Guieded Video Editing	MLLM Evaluation, Human Evaluation	BalanceCC [700]	0.4k
	TIP-Eval [683]	I2V Generation	DINO Similarity, DreamSim, CLIP-I, CLIP-T	TIP-I2V [683]	1700k
	AIGCBench [682]	I2V Generation	MSE, SSIM, CLIP-I, CLIP-T, RAFT, DOVER	WebVid-10M [395], LAION-5B [327]	1k
	VBench [667]	I2V Generation	DINO, MSE, CLIP-I, CLIP-T, RAFT, LAION	VBench [667]	2.2k
	VideoGen-Eval [691]	T2V Generation; I2V Generation	ViCLIP, Human Evaluation, MLLM Evaluation	VideoGen-Eval [691]	12k
	GenAI-Bench [692]	T2I Generation; T2V Generation	Human Evaluation, VQAScore	GenAI-Bench [692]	1.6k
Mix	InterleavedBench [693]	Interleaved Generation	MLLM Evaluation	WikiHow [376], ActivityNet Captions [701], MagicBrush [650], CustomDiffusion [702]	0.8k
	ISG-Bench [703]	Interleaved Generation	MLLM Evaluation	ISG-Bench [703]	1.1k
	UniEval [695]	Interleaved Generation	UniScore	UniEval [695]	4.2k
	SEED-Bench-2 [482]	Interleaved Generation	Accuracy	SEED-Bench-2 [482]	0.16k
	MMIE [694]	Interleaved Generation	MLLM Evaluation	WikiHow [376], VIST [357], MathVista [534], REMI [704]	20k
	CoMM [381]	T2I Generation; Interleaved Generation	FID, IS, SSIM, PSNR, MLLM Evaluation	MathVista [534], StoryGen [705]	1k
	MME-Unify [620]	Interleaved Generation; T2I ; Image Editing; T2V ; I2V	CLIP-I, CLIP-T, FID, FVD, Accuracy	COCO [312], ImagenHub [689], Emu-Edit [644], MSR-VTT [396], TIP-I2V [683], Pexel Video [706]	4.1k
	RealUnify [496]	Interleaved Generation	Accuracy, MLLM Evaluation	RealUnify [496], BLINK [484], HR-Bench [707]	1k

existing datasets to cover seven types of mathematical reasoning and five major task categories. MATHVERSE [537] provides high-quality problems in core mathematical fields and introduces a GPT-4-based Chain-of-Thought evaluation strategy for fine-grained analysis of reasoning steps.

To benchmark against human performance, recent datasets markedly increase difficulty. MATH-V [535] gathers thousands of problems from real-world mathematical competitions, categorized by sub-discipline and difficulty level. More over, OlympiadBench [536] curates Olympiad-level mathematics and physics problems complete with expert-annotated, step-by-step solutions. To investigate model robustness, MATH-P [539] confronts the "memorization versus reasoning" problem by applying controlled perturbations to challenging math problems.

OCR. Optical Character Recognition (OCR) benchmarks evaluate MLLMs' ability to recognize and understand text embedded in images. These tasks require models to understand the semantic context of recognized text to answer relevant questions, a capability crucial for interpreting information-dense media like webpages or diagrams, often termed rich-text understanding.

The evolution of these benchmarks shows a clear progression from basic recognition to sophisticated rich-text comprehension. Foundational benchmarks like TextVQA [515] provide large-scale datasets where answering questions necessitates reasoning about text present in images. WebSRC [432] extends this to webpages, introducing a structured reading comprehension task that demands a joint understanding of textual semantics, visual layout, and HTML structure. To facilitate more systematic evaluation, later work focuses on comprehensive coverage. For example, OCRBench [517] consolidates multiple text-related visual tasks, including recognition and VQA, into a unified benchmark. OCRBench v2 [518] then significantly expands this scope to cover numerous core reading skills across a wide range of tasks and real-world scenarios. Addressing specialized formats, SEED-Bench-2-Plus [519] features human-annotated multiple-choice questions to assess deep understanding of complex visuals like charts and maps.

Chart and Documentation. Charts and documents, characterized by structured layouts and rich textual content, require MLLMs to interpret graphical elements and understand relationships between visual and textual information.

Early benchmarks established the foundations for evaluating these capabilities. ChartQA [431] introduced a large-scale dataset of real-world charts accompanied by complex, human-authored reasoning questions. DocVQA [521] proposed a general-purpose document VQA task, emphasizing the need for models to synergistically leverage textual content and layout information. Focusing on a specific format, InfographicVQA [522] evaluates a model's capacity for holistic understanding of densely condensed visual and textual information. As MLLM capabilities advance, recent benchmarks increase the depth and difficulty of evaluation. MMLONGBENCH-DOC [524] targets long-document comprehension, evaluating critical skills like information localization and cross-page synthesis. In chart understanding, CharXiv [525] offers a realistic evaluation using complex charts from scientific papers and distinguishes between descriptive and inferential questions. Similarly, EvoChart-QA [532] uses diverse real-world charts to assess foundational comprehension skills. MMC-Benchmark [531] provides a fine-grained, nine-task framework that includes chart reasoning and conversion to structured formats like data tables. TableBench [533] orients towards industrial applications with a complex table VQA benchmark for tasks such as fact-checking and numerical reasoning.

Hallucination, defined as generated text that is inconsistent with or unsupported by visual input, is a major challenge in practical MLLM applications.

While early analyses adapted metrics like CHAIR [555], these methods have limited applicability to modern MLLMs. Consequently, POPE [546] introduces a probe-based methodology using binary "yes/no" questions to treat hallucination detection as a classification task, effectively revealing pervasive object-level inconsistencies. To capture more diverse errors, M-HalDetect [548] provides a dataset with fine-grained annotations that extend beyond object presence to include fabricated descriptions and inaccurate relationships. Recent methods leverage LLMs for evaluation and are designed to induce hallucinations. For example, HaELM [549] proposes a low-cost, LLM-based evaluation framework that assesses hallucinations within more realistic descriptive contexts. MMHAL-BENCH [550] is designed to penalize hallucinations, employing diverse questions to effectively induce and detect such behavior.

More recent works investigates the underlying causes of hallucination. For example, VLind-Bench [560] uses a series of prerequisite tests to disentangle failures stemming from language priors versus visual blind spots. Recently, VALOR-EVAL [557] expands assessment across object, attribute, and relationship dimensions, using challenging images to provoke hallucinations. It's a two-stage, LLM-based framework that jointly evaluates both the faithfulness and the informativeness of the generated text.

Safety and Bias. Safety and bias are essential evaluation criteria for contemporary MLLMs. Safety refers to a model's tendency to produce harmful content, while bias concerns outputs reflecting societal prejudices and stereotypes, such as those based on gender, age, or race.

Several benchmarks are designed to systematically assess these issues. MOSSBench [565] evaluates "multimodal hypersensitivity," where models over-conservatively refuse benign queries, revealing a fundamental tension between safety alignment and model utility. To address a broader spectrum of societal prejudices, VLBiasBench [566] utilizes large-scale synthetic data to evaluate MLLMs across numerous social and intersectional bias categories. Bingo [551] analyzes error sources in advanced MLLMs, dissecting phenomena like regional bias and visual-textual interference while evaluating mitigation strategies. Furthermore, MM-SPUBENCH [708] investigates "spurious bias" by constructing VQA tasks to quantify how models exploit non-essential correlations between input attributes and outputs.

9.1.2 Video

Foundation Capability. A series of evolving video question answering benchmarks assess the video understanding capabilities of MLLMs. Early works, such as TGIF-QA [580], pioneered this domain by focusing on counting, action recognition, and state transitions within short animated GIFs. As MLLMs advance, later benchmarks introduce more complex evaluations. For example, MMBench-Video [570] emphasizes temporal reasoning in long videos and leverages GPT-4 for robust, automated evaluation of free-form answers. Similarly, Video-MME [567] offers a comprehensive benchmark that integrates video, subtitle, and audio inputs to assess understanding across various video types.

Beyond general comprehension, recent benchmarks target more specialized skills. To assess temporal perception, TempCompass [572] uses "conflict videos" with identical static content but different temporal progressions, reducing reliance on static cues. MotionBench [576] and FAVOR-Bench [578] focus on fine-grained action understanding, with FAVOR-Bench incorporating egocentric videos and a novel LLM-free evaluation approach. VEU-Bench [573] evaluates video editing understanding, which requires high-level reasoning and domain knowledge.

To evaluate performance in complex interaction and reasoning scenarios, further benchmarks have been developed. V2P-Bench [575] uses visual prompts as an interactive modality to assess video understanding in precise human-computer collaboration. H²VU [577] provides a comprehensive evaluation for long-range reasoning, including tasks such as ultra-long video comprehension, counter-intuitive reasoning, and trajectory tracking. To address the high cost of data collection, VNBench [571] proposes a benchmark construction framework based on synthetic video generation, enabling efficient diagnosis of specific capabilities such as long-video retrieval and temporal ordering.

Long Context. Evaluating long-video comprehension for applications such as film analysis or embodied intelligence requires specialized benchmarks beyond those for short clips. Several works focus on long-context understanding. ActivityNet-QA [574] provides a large-scale, manually annotated dataset for assessing a model's understanding of fine-grained activities within extended video narratives. LVBench [588] measures long-term memory in videos spanning several hours through six core tasks that can be flexibly combined. HLV-1K [587] offers a benchmark of videos exceeding thirty minutes each, testing deep understanding

with hierarchical, time-aware questions. LongViTU [585] introduces a systematic approach for automated generation of long-video QA datasets.

To improve reliability and address specific contextual challenges, SVBench [582] introduces temporal multi-turn QA chains to evaluate a model's ability to leverage historical context in a streaming dialogue. To address shortcut bias in multiple-choice formats, CG-Bench [583] applies clue-based QA methods, while Event-Bench [589] uses an event-centric design to ensure genuine content understanding.

Recent benchmarks further extend to extreme-duration, first-person videos. EgoSchema [590] introduces an ultra-long egocentric QA dataset and the "temporal certificate set" to quantify temporal difficulty. X-LeBench [586] extends this with a life-logging simulation pipeline, generating egocentric videos up to 16.4 hours long, to assess long-term memory construction and retrieval.

World Knowledge. Recent efforts in video MLLM evaluation increasingly emphasize integrating external knowledge and domain-specific reasoning. Video-MMLU [591] focuses on understanding multi-disciplinary lecture videos, requiring complex reasoning over theorem demonstrations and problem-solving processes. Video SimpleQA [595] targets factual question answering, requiring integration of knowledge beyond the video's explicit content.

Benchmarks for real-world applications focus on specific types of intelligence. STI-Bench [593] measures spatiotemporal intelligence through tasks like predicting object pose, displacement, and motion. UrbanVideo-Bench [594] uses first-person drone videos from urban environments to evaluate perception and navigation in complex 3D settings.

Reasoning. Several benchmarks evaluate reasoning abilities in video understanding. For example, MINERVA [584] highlights interpretability by providing manually annotated, multi-step reasoning traces and introduces an error taxonomy to identify bottlenecks in visual perception and temporal localization. Additionally, InstructionBench [597] assesses temporal reasoning in instructional videos, using strict step-by-step procedures and filtering to ensure the evaluation focuses on visual analysis.

For advanced cognitive abilities, OVO-Bench [598] is an online video understanding benchmark that requires models to reason temporally by recalling historical context, understanding current events, and anticipating future information. VCBench [599] employs synthetic video generation for systematic evaluation of multiple cognitive abilities, especially in abstract or complex scenes.

OCR. Video OCR benchmarks systematically evaluate MLLMs' abilities in video-based optical character recognition and related comprehension tasks. FG-Bench [608] decomposes Video-OCR into six sub-tasks: text recognition, semantic understanding, spatial relationship reasoning, motion detection, text attribute identification, and temporal localization. To provide more comprehensive coverage and deeper evaluation, MME-VideoOCR [607] expands to 25 tasks across 10 categories and 44 diverse scenarios, emphasizing integration across frames, spatiotemporal reasoning, and mitigation of linguistic prior bias.

Hullucination, Safety and Bias. To ensure the reliability and safety of MLLMs in video understanding applications, researchers have developed a series of comprehensive evaluation benchmarks targeting the critical issues of hallucination, robustness, and bias.

VIDHAL [604] quantifies fine-grained hallucination levels in video MLLMs using a description-ranking task and novel ranking metrics, focusing on temporal hallucination. VideoHallucer [605] introduces a classification framework for hallucination and employs adversarial video question answering to probe model discrimination among different hallucination types. MVTamperBench [601] evaluates robustness against five common tampering methods (e.g., rotation, occlusion, replacement, repetition, frame dropping), unifying detection as a multiple-choice task to expose model vulnerabilities. VidLBEval [602] is the first benchmark to systematically evaluate language bias in video MLLMs, using ambiguous video comparisons and probing questions to measure over-reliance on textual over visual information.

9.1.3 Audio

To comprehensively evaluate the foundational capabilities of MLLMs in the domain of audio understanding, a series of targeted benchmarks have been constructed. MMAU [609] assesses expert-level knowledge and complex reasoning across speech, environmental sounds, and music, requiring models to exhibit numerous distinct skills. AIR-Bench [611] is a benchmark for instruction-following and open-ended generation in Large Audio-Language Models, using a hierarchical design and audio-mixing strategy with GPT-4 as a unified evaluator. AudioBench [612] offers broad coverage across many tasks and datasets to evaluate AudioLLMs on speech, audio scenes, and paralinguistic features, while also investigating the use of open-source models as evaluators.

In more specialized dimensions, SALMON [610] specifically evaluates a model's ability to comprehend non-semantic acoustic information within speech through tasks focused on acoustic consistency. For the music domain, MuChoMusic [613] is the first benchmark dedicated to music comprehension, using multiple-choice questions to probe knowledge of music theory and history. VoiceBench [614] evaluates the interactive performance of LLM-based voice assistants, assessing their knowledge, instruction-following, and safety across variations in speaker, environment, and content. To explore deeper logical reasoning, other benchmarks introduce novel tasks. Audio Entailment [615] requires a model to determine if a textual hypothesis can be inferred from an audio premise. Finally, Audiopedia [617] pioneers knowledge-intensive QA for the audio domain, structuring its evaluation into sub-tasks like single-audio, multi-audio, and retrieval-augmented QA.

9.1.4 Mix

Evaluating the ability of MLLMs to robustly and coherently process and integrate information from diverse modalities—text, image, audio, and video—is critical, as single-modality benchmarks are insufficient for assessing these complex conditions. Mixed-modality benchmarks are therefore essential for evaluating cross-modal reasoning and ensuring model reliability in real-world tasks.

To evaluate these complex capabilities, several benchmarks have been constructed. ACVUBench [618] is an audio-centric video understanding benchmark that uses audio-rich videos to evaluate a model's comprehension of

audio-video interactions and the role of audio in providing semantic cues. Advancing this, MAVERIX [619] evaluates the tight integration and joint reasoning over video and audio by introducing tasks that necessitate audio-video synergy for their resolution.

9.2 Benchmark for Generation

The development of Multimodal Large Language Models (MLLMs) has extended their capabilities beyond interpretation to sophisticated generative tasks. In addition to strong performance in understanding, recent models now demonstrate increasingly advanced generation across various modalities and application scenarios. This section surveys the benchmarks designed to evaluate these generative capabilities, covering image, video, and audio generation, as well as the synthesis of complex, mixed-modality outputs. We also discuss the evaluation dimensions and novel metrics proposed by these benchmarks to quantify the quality, coherence, and fidelity of generated content.

9.2.1 Image

Text-to-Image Generation. To systematically evaluate the rapidly advancing field of text-to-image (T2I) generation, researchers have developed a variety of benchmarks, each evaluating specific model abilities.

For compositional generation, benchmarks like T2I-CompBench [627], T2I-CompBench++ [632], and CONCEPTMIX [623] use structured prompts to assess a model's ability to generate multiple objects with complex attributes and relationships. PAINTSKILLS [635] uses synthetic data to evaluate fundamental visual reasoning skills, while DreamBooth [624] benchmarks personalized, subject-driven generation. To align evaluation with human perception, Pick-a-Pic [625] and its PickScore metric use user preference data to assess aesthetic quality. Other benchmarks focus on fidelity and world knowledge: TIFA [628] employs a Visual Question Answering (VQA) framework to measure alignment between generated images and text prompts, while WISE [641] and Commonsense-T2I [642] assess integration of world knowledge and commonsense reasoning using curated prompts and adversarial examples. I-HallA [634] uses a VQA-based approach to specifically assess image-level hallucinations and factual inaccuracies. Given potential societal impacts, several benchmarks address safety and fairness. FAIntbench [638] proposes a multi-dimensional system to define and measure fairness and bias, TIBET [709] uses counterfactual reasoning to quantify societal and incidental biases, and T2ISafety [639] introduces a hierarchical safety taxonomy and a dedicated evaluator, ImageGuard, to detect risks related to toxicity and privacy.

For comprehensive evaluation, platforms such as HEIM [640] integrate 12 dimensions—including alignment, aesthetics, and fairness—to standardize comparison of state-of-the-art models. TrustGen [636] offers a broader platform for evaluating generative model trustworthiness, dynamically assessing T2I models for authenticity and safety.

Image Editing. Instruction-driven image editing tasks require MLLMs to combine robust visual understanding with precise language comprehension to modify images according to user intent. This capability is evaluated by benchmarks focused on editing quality, instruction following, and preservation of non-edited content.

Several benchmarks address dataset quality and instruction complexity. To overcome dataset limitations, ULTRAEDIT [349] employs an automated pipeline to generate large-scale, high-quality editing samples, emphasizing real-world images and support for region-based operations. Emu Edit [644] introduces a multi-task learning framework with task embeddings for few-shot adaptation to complex instructions, providing a benchmark with seven editing operations. Focusing on fidelity and inversion accuracy, PIE-Bench [646] presents a diverse dataset of scenes and edit types, along with a comprehensive set of evaluation metrics.

Other benchmarks evaluate advanced reasoning in editing. RISEBench [651] measures temporal, causal, spatial, and logical inference within editing tasks using a comprehensive multi-dimensional evaluation framework. AURORA-BENCH [652] focuses on action- and reasoning-centric edits using minimally variant triplets from videos and simulations, also highlighting the limitations of automated evaluation metrics. ICEB [653] systematically evaluates large-scale concept editing, addressing generation issues caused by gaps in a model's internal knowledge.

9.2.2 Video

Text-to-Video Generation. Text-to-Video (T2V) generation require models to synthesize temporally coherent and semantically meaningful video from text prompts. Evaluating T2V models is essential to ensure that generated videos are visually convincing, logically consistent, and ethically safe.

Early benchmarks focus on video quality and alignment with human judgment. Frameworks such as EvalCrafter [656] and MiraBench [657] provide standardized prompts and metrics, while VIDEOFEEDBACK [665] leverages large-scale human feedback to create automated scores (VIDEOSCORE) that better reflect human preferences. For compositional evaluation, T2V-CompBench [627] systematically tests a model's ability to render scenes with specific attributes, spatial relationships, and action interactions.

Recent benchmarks further assess a model's adherence to real-world principles. PhyGenBench [672], Morpheus [674], and ChronoMagic-Bench2 [670] evaluate whether generated videos comply with physical laws and exhibit logical temporal progressions. StoryEval [675] specifically tests narrative coherence by verifying the sequence of events in generated stories.

As T2V models become more advanced, safety evaluation has become a priority. T2VSafetyBench [678] defines key safety dimensions, including unique temporal risks, and uses a dataset of adversarial prompts with GPT-4-assisted evaluation to identify vulnerabilities. This reflects a trend toward ensuring that generated videos are not only plausible but also logical and safe.

Text-Guided Video Editing. Text-guided video editing enables video modification via natural language instructions. This task requires models to balance faithful instruction following with temporal and content consistency. VE-Bench [679] introduces a large-scale dataset with human-annotated scores and a quantitative network to evaluate text–video alignment and source-to-edited consistency. EditBoard [680] offers a comprehensive evaluation framework

with nine automated metrics across four key dimensions (e.g., fidelity, consistency) and categorizes tasks by difficulty. In contrast, TGVE 2023 [681] provides an open-source collection of videos and prompts, relying on subjective human comparisons to evaluate text alignment, structural preservation, and aesthetics.

Image-to-Video Generation. Image-to-Video (I2V) generation transforms a static image into a dynamic video, often providing far greater control over initial content than T2V. Objective, user-aligned I2V benchmarks reflecting user perception are vital for progress. AIGCBench [682] provides a framework for fair algorithm comparison using four evaluation dimensions: control-video alignment, motion effect, temporal consistency, and video quality. TIP-I2V [683] introduces a large-scale dataset of over 1.7 million user prompts and corresponding videos, offering a valuable resource grounded in real-world user behavior. I2V-Bench [684] specifically evaluates consistency-enhancement techniques, emphasizing fine-grained assessment of visual consistency and dynamic logicality.

9.2.3 Audio.

Text & Speech to Speech. Text-to-Speech (TTS) and Speech-to-Speech (S2S) tasks include text-to-speech synthesis, voice conversion, and speech-to-speech translation. Evaluation focuses on clarity, naturalness, semantic accuracy, and the model's ability to follow instructions in interactive scenarios. EmphAssess [686] introduces an automated evaluation pipeline with a specialized test set and classifier to quantify how well models preserve emphasis across languages and speakers. TTSDS [687] proposes a multi-factor evaluation approach, decomposing speech quality into dimensions such as intelligibility, speaker characteristics, and environmental acoustics, using distribution-based comparisons between synthetic and real speech. This method is especially effective for measuring the naturalness and authenticity of S2S outputs. S2S-Arena [685] addresses S2S instruction-following in real-world tasks, constructing test scenarios from multiple domains and using arena-style pairwise comparisons to evaluate comprehensive performance, including paralinguistic features like emotion and intonation.

9.2.4 Mix

A number of benchmarks have been proposed to evaluate models on multiple generation tasks across modalities. For unified image generation and editing, ICE-Bench [688] introduces a systematic framework with a coarse-to-fine taxonomy covering 31 tasks and a suite of specialized automated metrics, including the advanced VLLM-QA. In contrast, ImagenHub [689] adopts a human-centric "gold standard" approach, using expert evaluation for seven key conditional generation tasks, such as mask-guided editing, subject-driven generation, and multi-concept composition.

Video generation benchmarks have also evolved. TC-Bench [690] focuses on temporal compositionality, evaluating both T2V and I2V generation using VLM-verified, assertion-based metrics. VideoGen-Eval proposes a dynamic, agent-based evaluation system, leveraging an LLM for prompt construction, an MLLM for content judgment, and specialized "patch tools" for T2V and I2V assessment. For both T2I and T2V tasks, GenAI-Bench [692] uses professionally curated prompts to identify weaknesses in complex compositional reasoning and supports improved generation quality through black-box re-ranking.

9.3 Benchmark for Mix Modality Generation.

The rapid development of UFMs, which integrate understanding and generation capabilities across multiple modalities, necessitates robust evaluation frameworks for mixed-modality generation. Current evaluations typically treat understanding and generation as separate tasks, employing disparate datasets and metrics that hinder meaningful cross-model comparisons. More critically, despite UFMs' potential for complex integrated tasks—such as instruction-guided image editing or visual reasoning with generative components—no standardized benchmark exists to systematically evaluate these unified capabilities.

To address these challenges, several benchmarks have emerged with distinct methodologies for evaluating mixed-modality generation. SEED-Bench-2 [482] employs large-scale multiple-choice questions for objective assessment of text-and-image generation, utilizing automated answer ranking for text and CLIP-based similarity for images to enable scalable evaluation. CoMM [381] constructs high-quality datasets through strict narrative filtering and multi-perspective alignment, supporting both traditional metrics (ROUGE, METEOR for text; FID, IS for images) and large-model-based scores from GPT-4o. InterleavedEval [693] adopts a reference-free approach, employing advanced MLLMs like GPT-4o as holistic judges across dimensions including text quality, perceptual coherence, and cross-modal alignment through rule-based scoring. ISG-BENCH [703] leverages scene graph structures to analyze fine-grained text-image relationships at multiple hierarchical levels, providing interpretable structural and content feedback. MMIE [694] compiles diverse academic queries and employs fine-tuned vision-language scoring models for bias-reduced, reliable evaluation. MME-Unify [620] standardizes evaluation across understanding (single/multi-image, video) and generation (image/video editing, T2V, I2V) tasks, introducing "Unify Tasks" requiring integrated reasoning and generation. UniEval [695] proposes a self-evaluation paradigm where models assess their own generated outputs through the UniScore metric, enabling unified, model-driven assessment. RealUnify [496] further evaluates whether UFMs can achieve genuine bidirectional synergy, in which understanding enhances generation and generation facilitates understanding, thereby enabling the solution of complex tasks; however, current architectures still fail to obtain tangible gains from such unification. These benchmarks collectively advance multimodal generation evaluation while differing in scale, granularity, data diversity, and reliance on human versus automated judgment.

10 Applications

With the development of UFMs, the notion of "Unify" has emerged as a prevailing paradigm beyond multimodal learning, extending to diverse domains such as robotics, autonomous driving, world models, medicine, and various

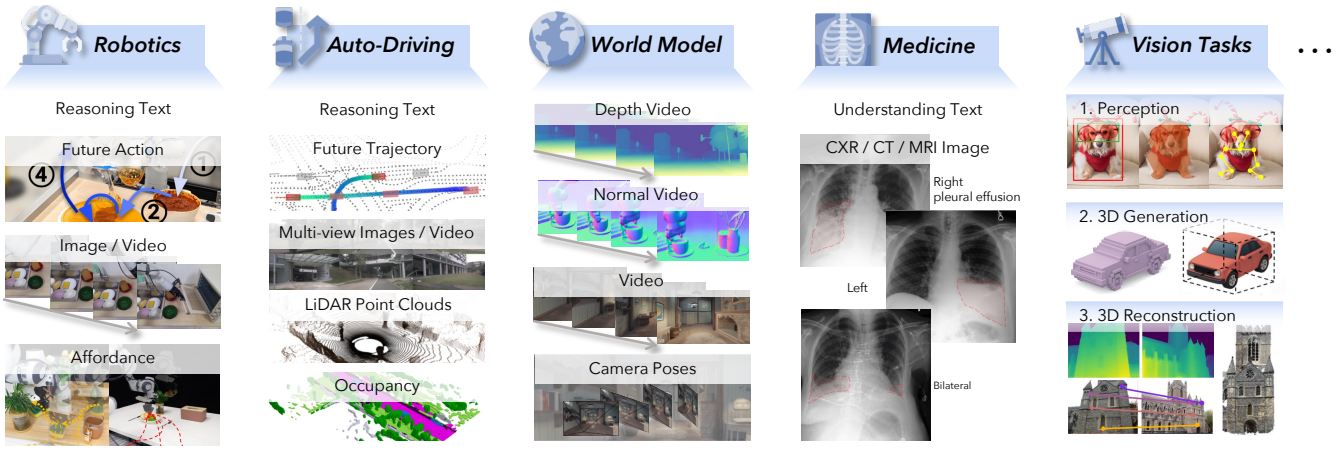

Fig. 17: Downstream applications of UFMs. We summarize the applications of UFMs in robotics, auto-driving, world models, medicine and vision tasks, detailed in Sec. 10. Representative output modalities of each domain are visualized.

vision tasks. Generating and predicting heterogeneous multimodal data allows models in these tasks to acquire the world knowledge inherent in large-scale datasets, thereby enhancing their scalability and generality.

10.1 Robotics

Robot policies based on vision-language-action (VLA) models generate future actions for planning from given observations and instructions [710]. For robotic systems such as manipulators and humanoid, planning poses unique challenges *e.g.*, handling multimodal action distributions [711]—which distinguish it from foundation tasks and often require specialized model structures, such as an additional diffusion head [712]. Recent advancements have shown great promise in integrating multimodal understanding and generation with action prediction.

To enhance model's action interpretability and reasoning, LCB [713], DexVLA [714], and ChatVLA [306] construct textual reasoning datasets to train models that explicitly provide analysis and reasoning before action denoising. CoA [715] further formalizes the reasoning process in embodied tasks, enabling efficient chain-of-affordances inference over objects, grasps, and spatial relations. To improve generalization, SEER [716] employs an auxiliary diffusion model to predict target action images, capturing real-world dynamics and agent–environment interactions, which offer dense visual supervision beyond text. Similarly, CoT-VLA [717] leverages the UFM [149] for model initialization to improve optimization efficiency, while DreamVLA [718] predicts depth maps and semantic knowledge, providing rich and compact representations. To exploit large-scale robotic video data without demonstrations, GR-2 [719] decouples video–action decoding and pretrains video generation on web-scale video data to capture general world dynamics. UVA [720] and UWM [721] further learns joint video–action latent representations from visual generative models [240], [722]. By selecting input–output modalities during training, these models enable versatile functionalities beyond policy learning, such as forward and inverse dynamics modeling, and video generation.

Overall, these advances show how unified robotic policies address key challenges of planning—interpretability, generalization, and scalability—paving the way for more versatile and robust embodied intelligence

10.2 Auto-Driving

Autonomous driving systems increasingly adopt end-to-end approaches that directly map raw sensor inputs to vehicle trajectories [723]. However, the complexity of 3D traffic environments typically necessitates auxiliary perception and prediction tasks with extensive annotations [724], [725]. Recent advances integrate future scene generation with motion planning, replacing traditional auxiliary tasks while enhancing dynamic environment modeling.

DrivingGPT [726] employs discrete autoregressive modeling to jointly predict interleaved future frames and motion tokens, leveraging implicit visual knowledge in driving scenes. Epona [727] decouples image and motion generation through combined autoregressive and diffusion models, enabling high-resolution, long-horizon prediction. UMGen [728] extends unified autoregression to additional modalities including raster maps and surrounding agents. Adriver-I [729] integrates future image prediction with driving VLA, while FSDrive [730] incorporates spatio-temporal physical priors by rendering lanes and detection boxes to guide VLA attention. Hermes [731] enhances 3D reasoning by predicting future LiDAR point clouds from current images across multimodal sensors. For fine-grained semantic understanding, OccLlama [732] and Occ-LLM [733] utilize spatial voxel representations from occupancy networks to predict future occupancy flow via driving VLA.

Collectively, these approaches highlight a paradigm shift toward unified generation and planning systems, enabling autonomous driving models to capture dynamic traffic environments with enhanced scene understanding, decision reliability, and generalization.

10.3 World Model

World models infer possible future states of the world from historical observations and alternative actions, and are regarded as one pathway toward autonomous intelligence [734]. However, world models that rely solely on 2D visual inputs often fail to maintain spatiotemporal coherence and adhere to physical constraints when generating

future dynamics, which limits their applicability to physics-oriented AI tasks such as data-driven simulation and robust policy learning [735]. Recent advancements have elevated world models through physically grounded, unified multi-modal prediction, enhancing their controllability and spatial reasoning capacity.

Since real-world scenes are fundamentally 3D rather than 2D pixel space, Aether [736] trains a diffusion-based 4D world model on synthetic 4D data, jointly predicting video, depth, and camera pose within a unified framework. This approach reconstructs 4D future observations, thereby embedding geometric priors into dynamic scene generation. DeepVerse [737] extends this line by retrieving spatial conditions from a global memory pool to reinforce temporal coherence in long-horizon 4D predictions. Unified world models have made encouraging strides in physical AI. TesserAct [738] constructs a 4D embodied world model that predicts not only video and depth but also normal maps, enabling more accurate capture of the scene's appearance, geometry, and surface properties. Such enriched representations enable the world model to simulate realistic 3D interactions, such as grasping objects or opening drawers. For driving tasks, GEM [739] trains a 4D driving world model with additional control conditions, enabling fine-grained operations in 4D futures, such as object manipulation and human pose changes. Further, DiST-4D [740] models multi-camera dependencies to generate consistent multi-view 4D futures, better reflecting the panoramic perception characteristics of real driving environments.

In summary, the developments of world models showcase the potential of unified world models for realistic, physically grounded integrative generation, providing a foundation for autonomous intelligence.

10.4 Medicine

Medicine represents one of the most promising application domains for multimodal intelligence [741]. Unified models, with their ability to jointly perform multimodal understanding and generation, have demonstrated significant potential across a wide range of medical tasks.

Recent advances demonstrate the feasibility of lightweight unified modeling in medicine. LLM-CXR [300] achieves efficient adaptation through instruction tuning of pretrained LLMs, eliminating costly domain-specific pretraining while supporting diverse tasks including CXR-to-report generation, report-to-CXR synthesis, and CXR-related VQA. Despite its lightweight design, LLM-CXR matches or exceeds specialized single-task models. MedXChat [303] integrates Stable Diffusion to enhance generative capabilities, achieving superior performance. HealthGPT [304] extends unified medical modeling beyond chest X-ray analysis to encompass OCT comprehension, microscopy image understanding, and CT-to-MRI generation, while investigating efficient training strategies for integrating heterogeneous medical tasks within a unified framework.

Nevertheless, deploying UFMs in medicine faces significant domain-specific challenges. Privacy and security constraints severely limit large-scale data collection, restricting model training and generalization capabilities. Additionally, the critical nature of healthcare demands exceptional reliability and safety standards, making hallucination and generation inconsistencies, which is already prevalent in current UFMs and particularly problematic. While UFMs demonstrate substantial potential for medical AI advancement, these constraints significantly impede their real-world clinical deployment.

10.5 Vision Tasks

Driven by the emergence of large-scale multimodal datasets and scalable learning techniques, vision models are transitioning from single-domain and single-task systems to versatile and general models [119]. Beyond generation and understanding tasks in UFMs, an increasing range of vision-centric tasks are now handled within a single unified framework to enhance the scalability and generalization.

Building upon the UFMs paradigm, recent advances have leveraged text and vision prompting to unify a wide range of vision perception tasks. LLMBind [189], Vision-LLM v2 [742], and Vitron [191] integrate task-specific vision models [743], [744], [745] as external decoders to handle high-level vision tasks, including object detection, semantic segmentation, and pose estimation. In contrast, UniWorld-V1 [200] and X-Prompt [218] leverage the generative capacity of UFMs to directly produce structured label figures, avoiding reliance on task-specific visual models. These methods also naturally extend to low-level vision tasks such as depth, normal, and edge estimation. Furthermore, Jodi [746] removes textual understanding, and jointly models the image domain and multiple label domains within a diffusion-based framework, enabling flexible switching between generation and perception tasks For 3D vision tasks, LLaMA-Mesh [302] leverages the OBJ format, which represents vertex coordinates and face definitions of 3D meshes as plain text, enabling direct integration with LLMs for 3D generation and understanding. ShapeLLM-Omni [747] trains a 3D VQVAE to convert 3D content into discrete tokens while preserving inherent topological structures for LLMs auto-regression. Furthermore, VGGT [748] serves as the first unified geometry foundation model that can directly infer all key 3D attributes of a scene from arbitrary views, including camera parameters, point maps, depth maps, and 3D point tracks. With its strong general 3D capability, it can reconstruct dense point clouds directly without any post-processing.

Overall, these advances show that the unified paradigm is emerging across distinct areas of vision,such as perception and 3D geometry, integrating diverse tasks within each domain to improve scalability and generalization.

11 Future Work and Discussion

11.1 Modeling Strategy and Structure

11.1.1 Autoregressive vs. Diffusion

With the rapid development of UFMs in recent years, both autoregressive and diffusion modeling have emerged as two main modeling paradigms, achieving significant progress and development, and leading to a surge of outstanding representative works [25], [145], [26], [226], [172], [225].

However, both traditional autoregressive modeling and diffusion modeling have their inherent flaws and limitations, which hinder further breakthroughs in UFM capabilities. For instance, the commonly used visual discretization [25], [145] in autoregressive modeling inevitably leads to information loss, while models [172] based on diffusion modeling still face significant shortcomings in terms of understanding capabilities. As a result, more and more recent works [117], [116], [22], [24] have adopted Autoregressive-Diffusion Hybrid Modeling. Currently, this strategy appears to better leverage the strengths of both modeling approaches, making it more suitable for modalities with distinct characteristics, such as text and images. For example, the incorporation of diffusion modeling has been shown to significantly improve the quality of image generation. Another modeling strategy that has gained increasing attention and adoption in recent years is integrating a lightweight diffusion head or flow head into autoregressive modeling. Inspired by MAR, these works combine the diffusion process and autoregressive backbone in a relatively decoupled manner, effectively avoiding the information loss caused by visual information discretization.

Another modeling strategy that has recently gained increasing attention and shows significant potential for future development is integrating a lightweight diffusion head or flow head into autoregressive modeling. Inspired by MAR, these approaches [219], [220], [161], [222], [749] combine the diffusion process and the autoregressive backbone in a relatively decoupled manner, effectively mitigating the information loss caused by visual discretization. Exclusively singular frameworks invariably possess unavoidable limitations. It is foreseeable that developing novel models by integrating the respective advantages of different frameworks, such as autoregressive and diffusion models, will constitute a highly influential research direction in the future.

11.1.2 Mixture of Experts

The MoE architecture has been widely applied and developed in the fields of LLMs and MLLMs, accumulating numerous valuable insights and experiences that contribute to the ongoing advancement of these models. Compared to dense models, MoE could achieve higher training efficiency with the same computational resources, and it also incurs lower inference costs with the same number of parameters. Additionally, MoE offers better scalability, enabling it to handle larger models and more complex tasks with ease.

In the field of UFMs, an increasing number of works [231], [22], [24], [217] are also adopting MoE or MoT architectures, reflecting the growing recognition of their benefits in improving model performance and efficiency. However, most of these works adopt a fixed-routing strategy, where different experts are assigned to different modalities, and modality interactions are facilitated through a shared self-attention layer. While this approach effectively learns modality-specific characteristics and helps avoid modality conflicts, relying solely on self-attention for interaction may limit deeper fusion and mutual enhancement of modality information. This issue warrants further exploration. Additionally, it is worth investigating whether further subdividing experts within the same modality could lead to more efficient and targeted processing.

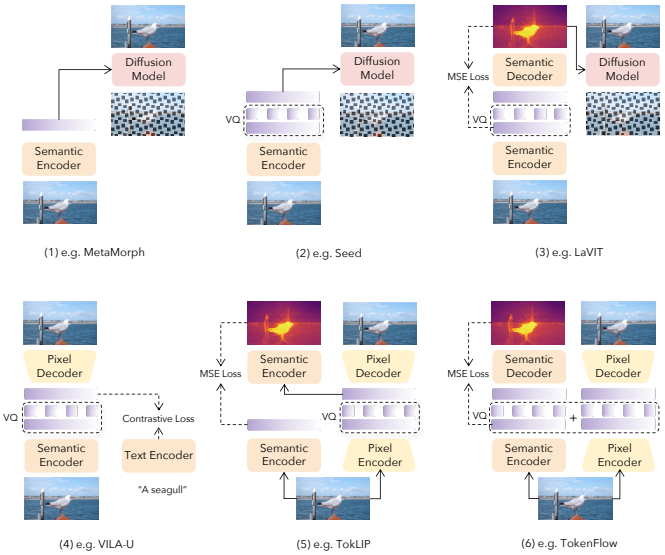

Fig. 18: Typical Unified Tokenizers of UFMs. Unified tokenizers can be classified into 6 typical categories. The encoder, latent representation, and the encoder are illustrated in each category, with annotations referring to classic unified tokenizer methods.

On the other hand, introducing MoE into UFMs inevitably brings challenges such as insufficient training stability, increased model complexity, and susceptibility to overfitting. These issues hinder the further development and application of MoE in this context. Therefore, how to effectively transfer the design and training experiences of MoE from LLMs and MLLMs to UFMs, while accumulating valuable insights specific to UFM modeling characteristics and developing a systematic theoretical framework, remains a significant challenge, but one that is valuable.

11.2 Unified Tokenzier

A tokenizer typically consists of an encoder and a decoder: the encoder transforms raw data into a compact latent space, while the decoder reconstructs the original data from these features. During training, the LLM only interacts with these features as input and output, and the decoder is only involved during inference. We refer to the tokenizer optimized for the UFM as the Unified Tokenizer. Equipped with the ability to capture both high-level semantic abstractions and fine-grained appearance details, it facilitates the UFM in unifying the learning paradigms of multimodal generation and understanding.

11.2.1 Tokenizer for Modular Joint Modeling

In modular joint modeling, encoders and decoders are typically constructed as independent models without joint alignment, as exemplified by MetaQueries [166], where a ViT encodes the input and a full diffusion model generates the output. As the encoder and decoder are not situated in a shared semantic space, the decoder need to participate in LLM's optimization through backpropagation. Although this enables high-quality generation by leveraging the full diffusion model, it inevitably introduces significant computational overhead.

Methods such as Emu2 [137] and MetaMorph [155] address this by pretraining diffusion models to condition on MLLM encoder inputs, as shown in Fig. 18 (1). The encoder is frozen to preserve the semantic feature space aligned with the LLM. This alignment establishes a Unified Tokenizer within a continuous semantic space. Training employs cosine similarity loss for continuous feature regression, with multimodal next token prediction [155] constraining the LLM to output complete image semantic features, enabling unified multimodal learning.

Regression in continuous feature space may lead to exponential error accumulation, particularly challenging for high-resolution images with larger token sequences. To address this, methods such as SEED [130] and DDT [204] introduce vector quantization to discretize the semantic space, as illustrated in Fig. 18 (2), thereby mitigating error propagation in continuous regression. These methods transform image features into 1D discrete token sequences rather than 2D spatial maps, employing techniques such as causal attention [130] or diffusion timestamp ordering [204]. This design enables LLMs to process multimodal data consistently with the original language modeling paradigm [130].

However, methods based on vector quantization and diffusion models cannot explicitly constrain the space of quantized features, which risks the loss of critical semantic information. To address this, approaches such as LaVIT [133] and ILLUME [164] introduce a semantic decoder to explicitly reconstruct the semantic feature maps aligned with the encoder, as illustrated in Fig. 18 (3). This design enforces semantic preservation in the quantized features while allowing the diffusion model to condition on more complete semantics, thereby enhancing generation quality.

11.2.2 Tokenizer for End-to-End Unified Modeling

End-to-end unified modeling typically employs general tokenizers [96], [282] directly, without requiring additional components like diffusion models. However, VQ-based tokenizers trained with reconstruction objectives often produce semantically impoverished features due to quantization constraints. Consequently, UFMs utilizing such tokenizers may exhibit degraded performance on understanding tasks.

Some methods integrate contrastive and reconstruction objectives to train unified tokenizers on image–text pairs. Some methods, such as VILA-U [149], QLIP [160], *et al.*, adopt a discretized high-level feature strategy, where a pretrained semantic encoder [31], [52] is used for initialization while retaining a pixel decoder. Building on the VQGAN [96] training paradigm, VILA-U performs contrastive learning between the quantized features and the corresponding text embeddings, thereby enhancing semantic information, as illustrated in fig. 18 (4). In contrast, TokLIP [170] follows a semanticized low-level feature strategy: while retaining the whole VQ tokenizer structure, the quantized features from a VQVAE encoder are further processed by a causal semantic encoder [31]. Supervised jointly with semantic feature distillation and contrastive learning, this design transforms low-level quantized features into semantic-aligned representations, as illustrated in fig. 18 (5). It is noteworthy that the latest RAE [18] employs the powerful feature extractors to reconstruct the VAE architecture, which has garnered significant attention and may potentially replace the functionality of most VAE modules.

Jointly training one encoder–decoder pair for both high-level semantic and low-level appearance features often leads to conflicting optimization objectives and instability. To address this, recent works adopt dual encoder–decoder architectures for decoupled feature learning, with different quantization fusion strategies, as illustrated in fig. 18 (6). TokenFlow [154] uses two encoder-specific codebooks with shared indices to maintain codebook size. SemHiTok [216] assigns an independent pixel codebook to each index in the semantic codebook, enhancing representational capacity but at the cost of a substantially larger all codebook size. MUSE-VL [215], in contrast, concatenates the two features before quantization, thus requiring only a single codebook. However, all these designs assume identical 2D spatial layouts for both encoders' features, which is oversimplified and inevitably introduces errors during the fusion stage.

11.2.3 Future Directions

Scalable Unified Tokenizer. In current end-to-end joint modeling, whether autoregressive or AR+Diffusion, low-level pixel space features are still predominantly used. This often necessitates an additional semantic encoder to enhance the UFM's understanding ability [22], [26]. Such a design is inelegant, introducing potential feature redundancy; moreover, in vision–language interleaving tasks, generated images need to be repeatedly re-encoded by the semantic encoder. Although many unified tokenizers have been proposed to extract "perfect features" that serve both generation and understanding, their design paradigms remain divergent and have yet to be validated at scale. Thus, the unified tokenizers with suitable architectures and trained on larger, higher-quality datasets will provide a solid foundation for the advancement of UFMs.

Efficient Video Tokenizer. As UFMs evolve, video understanding and generation will become increasingly important, with growing demands for longer duration and higher resolution. In existing end-to-end joint modeling, the number of tokens fed into the LLM scales linearly with the number of frames, posing severe challenges to context length and throughput. Some works alleviate this by reducing token counts through downsampling [117], [25], while others [175] directly employ a 3D VAE from a video generation model [277] for encoding and decoding video clips. In the future, unified video tokenizers with higher compression ratios and richer semantic representations will be of great value for advancing video-based UFMs.

Towards Omni-Modal Tokenzier. Since data from different modalities represent projections of a common underlying space, cross-modal representation learning may converge toward a shared representation as models scale [750]. This suggests the feasibility of constructing a single tokenizer that encodes and decodes diverse modalities (*e.g.*, text, image, audio, video) via a shared codebook in a unified latent space. Such an omni-modal tokenizer could enable more effective exploitation of large-scale cross-modal data, improve representation learning scalability, and facilitate modal-agnostic feature extraction. Furthermore, this tokenizer could provide unified interfaces across modalities

for the UFM, reducing alignment costs and enhancing multimodal generality (*e.g.*, generating audio-visual videos). Recent progress supports this vision: unified visual tokenizers for images and videos demonstrate notable advantages [751], [735], [752], while other efforts unify text, audio, and video [753], or incorporate 3D point clouds [754], into shared token spaces without decoders. Near-term advances in unified quantization schemes with effective modality-specific decoders appear most practical; long-term research toward universal encoder-decoder architectures with effective patching strategies remains an open frontier.

11.3 Model Training

The training paradigm for UFMs is expected to advance along two main trajectories. First, data construction will evolve from simple modality pairing to semantically interleaved corpora. Second, model optimization will move beyond conventional supervised fine-tuning toward unified frameworks that integrate reinforcement learning with human preference alignment.

11.3.1 Modality-Interleaved Data

The paradigm for constructing training data for unified models is shifting from simple modality pairing to deeply modality-interleaved corpora. Recent studies, such as AnyGPT [140], DreamLLM [132], and MIO [202], have widely adopted interleaved data for model training. Compared to models focused primarily on comprehension, unified models demand a higher degree of semantic consistency in such data. For instance, video data requires temporal decomposition at multiple granularities (e.g., frames, clips, events, and plots), and fine-grained alignment with text across dimensions like actions, objects, relationships, and causality. This strategy aims to simultaneously support both the comprehension and generation capabilities of the model, and its effectiveness has been validated in works like BAGEL [22]. Constructing such data significantly enhances the model's deep understanding of processes, behaviors, and complex temporal structures, enabling robust cross-modal alignment and reasoning in complex tasks.

However, constructing fine-grained, semantically interleaved data presents significant challenges. The core difficulty lies in the data construction process itself. On one hand, manual annotation is prohibitively expensive. Multi-granularity temporal decomposition and cross-modal semantic alignment require extensive expert labeling. Furthermore, subjective differences among annotators regarding semantic consistency can compromise data quality. On the other hand, while synthetic data generation can reduce costs, current controllable generation techniques face bottlenecks, making it difficult to precisely control the quality and diversity of the generated content. Given the reliance of large-scale models on massive datasets, developing high-precision controllable generation technologies [178] is considered a key research direction to address this bottleneck.

11.3.2 Training with Preference Alignment

In terms of training strategies, alignment fine-tuning based on reinforcement learning (e.g., DPO and GRPO) is another significant direction of development. Reinforcement learning strategies are increasingly being applied to the training of unified models, aiming to optimize output quality and align it with human preferences. For example, the DeepSeek-R1 [322] forgoes supervised fine-tuning and instead enhances its reasoning capabilities through large-scale reinforcement learning for post-training, demonstrating the immense potential of reinforcement learning-based training paradigms. Recently, Skywork UniPic2.0 [182] introduced the Progressive Dual-Task Reinforcement (PDTR) strategy for image generation and editing tasks, offering a new approach to alignment fine-tuning. These advancements indicate that preference alignment will play an increasingly central role in the future training of unified models.

However, applying preference alignment fine-tuning to unified models training also faces unique challenges. The core challenge lies in designing reward that can accurately capture human preferences across the dual tasks of understanding and generation. Traditionally, alignment fine-tuning has been primarily used to optimize generation tasks, while recent research has begun to explore preference alignment for dual-task scenarios [182]. It is foreseeable that future work will focus on designing algorithms capable of precisely capturing and balancing preferences for both understanding and generation tasks.

11.4 Benchmarks

We consider the unify capability of UFMs as understanding and generation that build on and support each other, rather than simply coexisting. The model should derive intermediate information that guide generation (e.g., plans or structured prompts), while the generative process should in turn provide visual evidence that strengthens understanding.

On the one hand, existing UFMs [23], [175] typically evaluate understanding and generation scores on corresponding mainstream benchmarks [476], [630], offering only indirect evaluation for unify capability. Although WISE's [641] T2I targets "understanding-to-generation" via world knowledge and complex semantics, there is still no direct assessment of generation-to-understanding or interleaved generation capability; some benchmarks [694], [693] evaluate text–image interleaving but do not explicitly measure mutual promotion. We suggest future unified benchmarks can (i) embed richer semantic and world-knowledge requirements into generation prompts so generation need to be grounded in a correct understanding of inputs, and (ii) require reasoning that depends on generated visuals (e.g., geometry problems needing constructed diagrams), thereby testing how generation facilitates understanding.

On the other hand, current metrics can be categorized into (a) accuracy for discrete answers (e.g., SEED-Bench-v2 [482], UniEval [695], MME-Unify [620]), which obscures whether intermediate reasoning and visual evidence are valid, and (b) MLLM-as-Judge (e.g., MMIE [694], WISE [641]), which improves process transparency but introduces bias from the judge model's data and scale. To this end, we may propose a hybrid protocol that use an MLLM-as-Judge to evaluate intermediate output (e.g., whether plans improved instruction-following or whether generated visuals aided downstream reasoning), while assessing final outputs with objective, rule-or reference-based measures.

This combination preserves interpretability of mutual promotion and yields robust, comparable end metrics.

12 Conclusion

This paper presents a comprehensive survey of UFMs that integrate multimodal understanding and generation capabilities within a single framework. Through a systematic analysis of over 700 publications, we examine the technological evolution, methodological taxonomies, prevailing challenges, and practical applications in this rapidly advancing field through 11 chapters. Despite significant progress, several critical challenges remain unresolved, including suboptimal multimodal alignment mechanisms, inherent conflicts between understanding and generation objectives, scarcity of high-quality unified training data, absence of standardized evaluation protocols, etc. We discuss promising future research directions encompassing UFM architectures, unified tokenzier design, human preference alignment, and standardized benchmarking frameworks. This survey contributes to the field by providing a structured foundation for addressing existing technical limitations in UFMs, facilitating progress toward artificial general intelligence, and offering actionable insights for the research community.

References

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin , *et al.*, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR)*, 2019. 4

[2] M. Elasri, O. Elharrouss, S. Al-Maadeed , *et al.*, "Image generation: A review," *Neural Processing Letters*, 2022. 4

[3] H. Liu, C. Li, Y. Li , *et al.*, "Improved baselines with visual instruction tuning," in *CVPR*, 2024. 4

[4] S. Bai, K. Chen, X. Liu , *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025. 4

[5] K. K. Team, B. Yang, B. Wen , *et al.*, "Kwai keye-vl technical report," *arXiv preprint arXiv:2507.01949*, 2025. 4

[6] L.-C.-T. Xiaomi, "Mimo-vl technical report," *arXiv preprint arXiv:2506.03569*, 2025. 4

[7] B. S. Team, "Seed1.5-vl technical report," *arXiv preprint arXiv:2505.07062*, 2025. 4

[8] Z. Chen, W. Wang, H. Tian , *et al.*, "Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy," 2024. [Online]. Available: https://internvl.github.io/blog/2024-07-02-InternVL-2.0 4, 7

[9] C. Fu, H. Lin, X. Wang , *et al.*, "Vita-1.5: Towards gpt-4o level real-time vision and speech interaction," *arXiv preprint arXiv:2501.01957*, 2025. 4

[10] OpenAI, "Hello gpt-4o," 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/ 4, 15, 16, 53

[11] Google Developers, "Experiment with gemini 2.0 flash native image generation," *Google Developers Blog*, 2024. 4, 51

[12] B. Yang, B. Wen, B. Ding , *et al.*, "Kwai keye-vl 1.5 technical report," *arXiv preprint arXiv:2509.01563*, 2025. 4

[13] F.-A. Croitoru, V. Hondru, R. T. Ionescu , *et al.*, "Diffusion models in vision: A survey," *IEEE TPAMI*, 2023. 4

[14] B. F. Labs, "Flux," 2024. [Online]. Available: https://github.com/black-forest-labs/flux 4, 30, 31, 32

[15] R. Rombach, A. Blattmann, D. Lorenz , *et al.*, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022. 4, 7, 8, 14, 15, 30, 31, 32, 38

[16] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023. 4, 8, 22, 36

[17] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023. 4, 20

[18] B. Zheng, N. Ma, S. Tong , *et al.*, "Diffusion transformers with representation autoencoders," *arXiv preprint arXiv:2510.11690*, 2025. 4, 8, 67

[19] OpenAI, "Thinking with images," 2025. [Online]. Available: https://openai.com/index/thinking-with-images/ 4

[20] H. Zhang, C. Li, W. Wu , *et al.*, "Scaling and beyond: Advancing spatial reasoning in mllms requires new recipes," *arXiv preprint arXiv:2504.15037*, 2025. 4

[21] Y.-F. Zhang, X. Lu, S. Yin , *et al.*, "Thyme: Think beyond images," *arXiv preprint arXiv:2508.11630*, 2025. 4

[22] C. Deng, D. Zhu, K. Li , *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025. 4, 11, 18, 22, 24, 29, 32, 33, 37, 41, 48, 51, 53, 66, 67, 68

[23] J. Chen, Z. Xu, X. Pan , *et al.*, "Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset," *arXiv preprint arXiv:2505.09568*, 2025. 4, 11, 15, 24, 30, 31, 36, 37, 39, 47, 55, 68

[24] C. Liao, L. Liu, X. Wang , *et al.*, "Mogao: An omni foundation model for interleaved multi-modal generation," *arXiv preprint arXiv:2505.05472*, 2025. 4, 18, 22, 24, 29, 32, 33, 37, 38, 41, 43, 44, 66

[25] X. Wang, X. Zhang, Z. Luo , *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024. 4, 11, 17, 18, 19, 24, 26, 27, 30, 34, 35, 36, 37, 40, 47, 49, 65, 66, 67

[26] X. Chen, Z. Wu, X. Liu , *et al.*, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025. 4, 11, 17, 18, 19, 37, 38, 41, 47, 48, 65, 67

[27] D. Podell, Z. English, K. Lacey , *et al.*, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023. 4, 8, 30, 31, 32

[28] K. He, X. Zhang, S. Ren , *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016. 5

[29] K. He, H. Fan, Y. Wu , *et al.*, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020. 5

[30] T. Chen, S. Kornblith, M. Norouzi , *et al.*, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020. 5

[31] A. Radford, J. W. Kim, C. Hallacy , *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021. 5, 15, 16, 18, 24, 31, 32, 34, 36, 51, 67

[32] C. Jia, Y. Yang, Y. Xia , *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021. 5

[33] X. Zhai, B. Mustafa, A. Kolesnikov , *et al.*, "Sigmoid loss for language image pre-training," *arXiv preprint arXiv:2303.15343*, 2023. 5, 15, 18, 23, 24, 25, 31, 32, 36

[34] A. Kolesnikov, L. Beyer, X. Zhai , *et al.*, "Big transfer (bit): General visual representation learning," in *ECCV*, 2020. 5, 6

[35] H. Pham, Z. Dai, G. Ghiasi , *et al.*, "Combined scaling for zero-shot transfer learning," *Neurocomputing*, 2023. 5

[36] L. Yuan, D. Chen, Y.-L. Chen , *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021. 5

[37] M. Cherti, R. Beaumont, R. Wightman , *et al.*, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023. 5

[38] B. Zhu, B. Lin, M. Ning , *et al.*, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," *arXiv preprint arXiv:2310.01852*, 2023. 5

[39] R. Girdhar, A. El-Nouby, Z. Liu , *et al.*, "Imagebind: One embedding space to bind them all," in *CVPR*, 2023. 5, 24, 32

[40] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021. 6

[41] H. Zhang, F. Li, S. Liu , *et al.*, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022. 6

[42] M. Assran, M. Caron, I. Misra , *et al.*, "Masked siamese networks for label-efficient learning," in *ECCV*, 2022. 6

[43] C. Sun, A. Shrivastava, S. Singh , *et al.*, "Revisiting unreasonable effectiveness of data in deep learning era," in *ICCV*, 2017. 6

[44] W. Wu, A. Timofeev, C. Chen , *et al.*, "Mofi: Learning image representations from noisy entity annotated images," *arXiv preprint arXiv:2306.07952*, 2023. 6

[45] J. Devlin, M.-W. Chang, K. Lee , *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL (long and short papers)*, 2019. 7

[46] H. Bao, L. Dong, S. Piao , *et al.*, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021. 7

[47] R. Wang, D. Chen, Z. Wu , *et al.*, "Bevt: Bert pretraining of video transformers," in *CVPR*, 2022. 7

[48] K. He, X. Chen, S. Xie , *et al.*, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022. 7

[49] Z. Tong, Y. Song, J. Wang , *et al.*, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *NeurIPS*, 2022. 7

[50] Z. Xie, Z. Zhang, Y. Cao , *et al.*, "Simmim: A simple framework for masked image modeling," in *CVPR*, 2022. 7

[51] Y. Fang, W. Wang, B. Xie , *et al.*, "Eva: Exploring the limits of masked visual representation learning at scale," in *CVPR*, 2023. 7

[52] Y. Fang, Q. Sun, X. Wang , *et al.*, "Eva-02: A visual representation for neon genesis," *Image and Vision Computing*, 2024. 7, 67

[53] Q. Sun, Y. Fang, L. Wu , *et al.*, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023. 7, 15, 24, 25, 31

[54] A. Radford, J. Wu, R. Child , *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019. 7

[55] H. Liu, C. Li, Q. Wu , *et al.*, "Visual instruction tuning," *NeurIPS*, 2023. 7, 51, 53, 55, 56

[56] H. Liu, C. Li, Y. Li , *et al.*, "Improved baselines with visual instruction tuning," in *CVPR*, 2024. 7, 46, 48, 55, 56

[57] J. Bai, S. Bai, S. Yang , *et al.*, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023. 7, 31

[58] H. Lu, W. Liu, B. Zhang , *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024. 7

[59] P. Wang, S. Bai, S. Tan , *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. 7, 31

[60] Y. Chu, J. Xu, Q. Yang , *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024. 7

[61] J. Xu, Z. Guo, J. He , *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025. 7

[62] Z. Liu, Y. Dong, Z. Liu , *et al.*, "Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution," *arXiv preprint arXiv:2409.12961*, 2024. 7

[63] Z. Liu, Y. Dong, J. Wang , *et al.*, "Ola: Pushing the frontiers of omni-modal language model," *arXiv preprint arXiv:2502.04328*, 2025. 7

[64] Y. Dong, Z. Liu, H.-L. Sun , *et al.*, "Insight-v: Exploring long-chain visual reasoning with multimodal large language models," in *CVPR*, 2025, pp. 9062–9072. 7

[65] Y. Shi, J. Liu, Y. Guan , *et al.*, "Mavors: Multi-granularity video representation for multimodal large language model," *arXiv preprint arXiv:2504.10068*, 2025. 7

[66] Y. W. Teh, M. Welling, S. Osindero , *et al.*, "Energy-based models for sparse overcomplete representations," *Journal of Machine Learning Research*, 2003. 7

[67] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza , *et al.*, "Generative adversarial nets," *NeurIPS*, 2014. 7

[68] U. Michelucci, "An introduction to autoencoders," *arXiv preprint arXiv:2201.03898*, 2022. 7, 8

[69] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu , *et al.*, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2023. 7, 8, 9, 31, 32

[70] Y. LeCun, S. Chopra, R. Hadsell , *et al.*, "A tutorial on energy-based learning," *Predicting structured data*, 2006. 7

[71] J. Xie, S.-C. Zhu, and Y. Nian Wu, "Synthesizing dynamic patterns by spatial-temporal generative convnet," in *CVPR*, 2017. 7

[72] D. Yilun and I. Mordatch, "Implicit generation and generalization in energy-based models," *arXiv preprint arXiv:1903.08689*, 2020. 7

[73] E. Nijkamp, M. Hill, S.-C. Zhu , *et al.*, "Learning non-convergent non-persistent short-run mcmc toward energy-based model," *NeurIPS*, 2019. 7

[74] G. Raut and A. Singh, "Generative ai in vision: A survey on models, metrics and applications," *arXiv preprint arXiv:2402.16369*, 2024. 7, 8, 9

[75] S. Bengesi, H. El-Sayed, M. K. Sarker , *et al.*, "Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers," *IEEE Access*, 2024. 7, 8

[76] C. Ledig, L. Theis, F. Huszár , *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017. 7

[77] S. Vasu, N. Thekke Madam, and A. Rajagopalan, "Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network," in *Proceedings of the ECCV (ECCV) Workshops*, 2018. 7

[78] P. Isola, J.-Y. Zhu, T. Zhou , *et al.*, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017. 7, 8

[79] A. Brock, T. Lim, J. Ritchie , *et al.*, "Neural photo editing with introspective adversarial networks," in *ICLR*, 2017. 7

[80] R. Huang, S. Zhang, T. Li , *et al.*, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *ICCV*, 2017. 7

[81] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019. 7

[82] G. Iglesias, E. Talavera, and A. Díaz-Álvarez, "A survey on gans for computer vision: Recent research, analysis and taxonomy," *Computer Science Review*, 2023. 7

[83] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, 2019. 7

[84] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 7

[85] T. Miyato, T. Kataoka, M. Koyama , *et al.*, "Spectral normalization for generative adversarial networks," in *ICLR*, 2018. 7

[86] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017. 7

[87] I. Gulrajani, F. Ahmed, M. Arjovsky , *et al.*, "Improved training of wasserstein gans," *NeurIPS*, 2017. 7

[88] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," in *ICLR*, 2017. 7

[89] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017. 7

[90] H. Zhang, I. Goodfellow, D. Metaxas , *et al.*, "Self-attention generative adversarial networks," in *ICML*. PMLR, 2019. 7

[91] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *ICLR*, 2017. 7

[92] D. E. Rumelhart, J. L. McClelland, P. R. Group , *et al.*, *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986. 8

[93] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014. 8, 23, 24

[94] A. Van Den Oord, O. Vinyals , *et al.*, "Neural discrete representation learning," *NeurIPS*, 2017. 8, 24, 26, 30

[95] A. Van den Oord, N. Kalchbrenner, L. Espeholt , *et al.*, "Conditional image generation with pixelcnn decoders," *NeurIPS*, 2016. 8, 9

[96] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021. 8, 17, 18, 24, 26, 27, 30, 32, 34, 38, 39, 67

[97] J. Lu, C. Clark, R. Zellers , *et al.*, "Unified-io: A unified model for vision, language, and multi-modal tasks," in *ICLR*, 2023. 8, 11, 18, 22, 23, 26, 34, 36, 37, 38, 40, 43, 47

[98] E. Aiello, Y. Lili, Y. Nie , *et al.*, "Jointly training large autoregressive multimodal models," in *ICLR*, 2023. 8, 18, 34, 37, 44

[99] A. Nichol, P. Dhariwal, A. Ramesh , *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021. 8

[100] C. Saharia, W. Chan, S. Saxena , *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, 2022. 8, 58, 59

[101] Y. Song, P. Dhariwal, M. Chen , *et al.*, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023. 8, 34

[102] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015. 8, 32

[103] P. Esser, S. Kulal, A. Blattmann , *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first ICML*, 2024. 9, 20, 30, 31, 32

[104] N. Ma, M. Goldstein, M. S. Albergo , *et al.*, "Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers," in *ECCV*, 2024. 9

[105] T. Brooks, B. Peebles, C. Holmes , *et al.*, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators 9

[106] G. Papamakarios, E. Nalisnick, D. J. Rezende , *et al.*, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, 2021. 9

[107] W. Kaplan, *Advanced calculus*. Pearson Education India, 1952. 9
[108] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014. 9
[109] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," in *ICLR*, 2017. 9
[110] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *NeurIPS*, 2018. 9
[111] D. Tran, K. Vafa, K. Agrawal , *et al.*, "Discrete flows: Invertible generative models of discrete data," *NeurIPS*, 2019. 9
[112] A. Oord, Y. Li, I. Babuschkin , *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *ICML*, 2018. 9
[113] J. Xiong, G. Liu, L. Huang , *et al.*, "Autoregressive models in vision: A survey," *Transactions on Machine Learning Research*, 2025. 9
[114] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *ICML*, 2016. 9
[115] K. Tian, Y. Jiang, Z. Yuan , *et al.*, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *NeurIPS*, 2024. 9, 19, 27, 30, 34
[116] J. Xie, W. Mao, Z. Bai , *et al.*, "Show-o: One single transformer to unify multimodal understanding and generation," in *ICLR*, 2025. 9, 11, 18, 21, 22, 24, 34, 35, 37, 39, 40, 44, 45, 48, 66
[117] C. Zhou, L. YU, A. Babu , *et al.*, "Transfusion: Predict the next token and diffuse images with one multi-modal model," in *ICLR*, 2025. 9, 18, 19, 21, 22, 24, 30, 32, 37, 44, 66, 67
[118] H. Chung, D. Lee, and J. C. Ye, "Acdc: Autoregressive coherent multimodal generation using diffusion correction," *arXiv preprint arXiv:2410.04721*, 2024. 9
[119] C. Fu, H. Lin, Z. Long , *et al.*, "Vita: Towards open-source interactive omni multimodal llm," *arXiv preprint arXiv:2408.05211*, 2024. 10, 65
[120] R. Huang, M. Li, D. Yang , *et al.*, "Audiogpt: Understanding and generating speech, music, sound, and talking head," in *AAAI*, 2024. 10, 12, 37
[121] D. Zhang, S. Li, X. Zhang , *et al.*, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," in *EMNLP*, 2023. 10, 17, 18
[122] S. Jiang, J. Liang, J. Wang , *et al.*, "From specific-mllms to omni-mllms: A survey on mllms aligned with multi-modalities," *arXiv preprint arXiv:2412.11694*, 2024. 10
[123] P. Wang, A. Yang, R. Men , *et al.*, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *ICML*, 2022. 11, 18, 22, 23, 34, 37, 39, 43, 47, 48
[124] X. Xu, Z. Wang, E. Zhang , *et al.*, "Versatile diffusion: Text, images and variations all in one diffusion model." in *ICCV*, 2023. 11, 18, 20
[125] C. Wu, S. Yin, W. Qi , *et al.*, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023. 10, 11, 12, 37
[126] Y. Shen, K. Song, X. Tan , *et al.*, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *NeurIPS*, 2023. 10, 11, 12, 13, 37
[127] F. Bao, S. Nie, K. Xue , *et al.*, "One transformer fits all distributions in multi-modal diffusion at scale," in *ICML*, 2023. 11, 18, 20
[128] Z. Tang, Z. Yang, C. Zhu , *et al.*, "Any-to-any generation via composable diffusion," *NeurIPS*, 2023. 11, 16, 18, 20, 33, 37, 41, 43, 45
[129] Q. Sun, Q. Yu, Y. Cui , *et al.*, "Emu: Generative pretraining in multimodality," in *ICLR*, 2024. 11, 14, 17, 29, 37, 38, 39, 43
[130] Y. Ge, Y. Ge, Z. Zeng , *et al.*, "Planting a seed of vision in large language model," *arXiv preprint arXiv:2307.08041*, 2023. 11, 15, 24, 27, 28, 30, 35, 37, 40, 41, 67
[131] S. Wu, H. Fei, L. Qu , *et al.*, "Next-gpt: Any-to-any multimodal llm," in *ICML*, 2024. 11, 14, 16, 29, 30, 31, 32, 37, 39, 41, 48
[132] R. Dong, C. Han, Y. Peng , *et al.*, "Dreamllm: Synergistic multimodal comprehension and creation," in *ICLR*, 2024. 11, 14, 16, 29, 31, 37, 68
[133] Y. Jin, K. Xu, L. Chen , *et al.*, "Unified language-vision pretraining in llm with dynamic discrete visual tokenization," in *ICLR*, 2024. 11, 15, 24, 30, 35, 36, 37, 38, 39, 43, 67
[134] K. Zheng, X. He, and X. E. Wang, "Minigpt-5: Interleaved vision-and-language generation via generative vokens," *arXiv preprint arXiv:2310.02239*, 2023. 11, 14, 24, 30, 31, 37, 42, 44, 49
[135] Y. Ge, S. Zhao, Z. Zeng , *et al.*, "Making llama see and draw with seed tokenizer," in *ICLR*, 2024. 11, 15, 25, 35, 37
[136] Z. Tang, Z. Yang, M. Khademi , *et al.*, "Codi-2: In-context interleaved and interactive any-to-any generation," in *CVPR*, 2024. 11, 16, 30, 31, 32, 37, 39, 49
[137] Q. Sun, Y. Cui, X. Zhang , *et al.*, "Generative multimodal models are in-context learners," in *CVPR*, 2024. 11, 14, 15, 17, 24, 25, 30, 31, 32, 37, 48, 67
[138] J. Lu, C. Clark, S. Lee , *et al.*, "Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action," in *CVPR*, 2024. 11, 18, 22, 23, 34, 35, 37, 38, 41, 43, 47, 48
[139] C. Tian, X. Zhu, Y. Xiong , *et al.*, "Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer," *arXiv preprint arXiv:2401.10208*, 2024. 11, 14, 24, 29, 30, 32, 37, 44, 48
[140] J. Zhan, J. Dai, J. Ye , *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," in *ACL (1)*, 2024. 11, 14, 16, 27, 35, 37, 40, 43, 48, 68
[141] Y. Jin, Z. Sun, K. Xu , *et al.*, "Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization," in *ICML*, 2024. 11, 15, 24, 30, 35, 37
[142] H. Liu, W. Yan, M. Zaharia , *et al.*, "World model on million-length video and language with blockwise ringattention," in *ICLR*, 2024. 11, 17, 18, 26, 27, 34, 35, 37
[143] Y. Li, Y. Zhang, C. Wang , *et al.*, "Mini-gemini: Mining the potential of multi-modality vision language models," *arXiv preprint arXiv:2403.18814*, 2024. 11, 13, 37
[144] Y. Ge, S. Zhao, J. Zhu , *et al.*, "Seed-x: Multimodal models with unified multi-granularity comprehension and generation," *arXiv preprint arXiv:2404.14396*, 2024. 11, 29, 30, 31, 32, 37, 39
[145] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," *arXiv preprint arXiv:2405.09818*, 2024. 11, 18, 19, 34, 37, 38, 44, 48, 65, 66
[146] Z. Zhao, J. Tang, B. Wu , *et al.*, "Harmonizing visual text comprehension and generation," *NeurIPS*, 2024. 11, 15, 37, 41
[147] E. Chern, J. Su, Y. Ma , *et al.*, "Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation," *arXiv preprint arXiv:2407.06135*, 2024. 11, 18, 19, 37, 43, 44
[148] D. Liu, S. Zhao, L. Zhuo , *et al.*, "Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining," *arXiv preprint arXiv:2408.02657*, 2024. 11, 18, 19, 34
[149] Y. Wu, Z. Zhang, J. Chen , *et al.*, "Vila-u: a unified foundation model integrating visual understanding and generation," *arXiv preprint arXiv:2409.04429*, 2024. 11, 18, 24, 26, 28, 30, 34, 35, 37, 38, 41, 44, 47, 64, 67
[150] C. Wu, X. Chen, Z. Wu , *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," in *CVPR*, 2025. 11, 15, 17, 18, 19, 22, 28, 29, 34, 37, 40
[151] R. Fang, C. Duan, K. Wang , *et al.*, "Puma: Empowering unified mllm with multi-granular visual generation," *arXiv preprint arXiv:2410.13861*, 2024. 11, 15, 24, 29, 30, 31, 32, 37, 38
[152] Y. Ma, X. Liu, X. Chen , *et al.*, "Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation," in *CVPR*, 2025. 11, 18, 22, 24, 29, 30, 32, 33, 37
[153] W. Liang, L. YU, L. Luo , *et al.*, "Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models," in *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025. 11, 37, 44
[154] L. Qu, H. Zhang, Y. Liu , *et al.*, "Tokenflow: Unified image tokenizer for multimodal understanding and generation," in *CVPR*, 2025. 11, 18, 19, 24, 27, 28, 30, 34, 36, 37, 40, 41, 67
[155] S. Tong, D. Fan, J. Zhu , *et al.*, "Metamorph: Multimodal understanding and generation via instruction tuning," *arXiv preprint arXiv:2412.14164*, 2024. 11, 15, 24, 25, 30, 31, 37, 67
[156] J. Wu, Y. Jiang, C. Ma , *et al.*, "Liquid: Language models are scalable and unified multi-modal generators," *arXiv preprint arXiv:2412.04332*, 2024. 11, 17, 18, 37
[157] S. Liu, A. S. Hussain, Q. Wu , *et al.*, "Mumu-llama: Multi-modal music understanding and generation via large language models," *arXiv preprint arXiv:2412.06660*, 2024. 11, 16, 26, 30, 32, 37
[158] X. Zhuang, Y. Xie, Y. Deng , *et al.*, "Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model," *arXiv preprint arXiv:2501.12327*, 2025. 11, 18, 19, 34, 37, 47
[159] C. Ma, Y. Jiang, J. Wu , *et al.*, "Unitok: A unified tokenizer for visual generation and understanding," *arXiv preprint arXiv:2502.20321*, 2025. 11, 18, 19, 24, 28, 30, 34, 37, 44

[160] Y. Zhao, F. Xue, S. Reed , *et al.*, "Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation," *arXiv preprint arXiv:2502.05178*, 2025. 11, 18, 24, 30, 34, 37, 67

[161] S. Wu, W. Zhang, L. Xu , *et al.*, "Harmonizing visual representations for unified multimodal understanding and generation," *arXiv preprint arXiv:2503.21979*, 2025. 11, 18, 19, 33, 37, 66

[162] A. Swerdlow, M. Prabhudesai, S. Gandhi , *et al.*, "Unified multimodal discrete diffusion," *arXiv preprint arXiv:2503.20853*, 2025. 11, 18, 20, 34, 37, 39, 45

[163] J. Zou, B. Liao, Q. Zhang , *et al.*, "Omnimamba: Efficient and unified multimodal understanding and generation via state space models," *arXiv preprint arXiv:2503.08686*, 2025. 11, 18, 22, 23, 24, 34, 37

[164] C. Wang, G. Lu, J. Yang , *et al.*, "Illume: Illuminating your llms to see, draw, and self-enhance," *ICCV*, 2025. 11, 15, 24, 28, 30, 35, 37, 67

[165] Y. Jiao, H. Qiu, Z. Jie , *et al.*, "Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding," in *CVPR*, 2025. 11, 17, 18, 24, 28, 37

[166] X. Pan, S. N. Shukla, A. Singh , *et al.*, "Transfer between modalities with metaqueries," *arXiv preprint arXiv:2504.06256*, 2025. 11, 16, 29, 30, 31, 37, 44, 51, 66

[167] X. Zhuang, Y. Xie, Y. Deng , *et al.*, "Vargpt-v1. 1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning," *arXiv preprint arXiv:2504.02949*, 2025. 11, 18, 19, 29, 34, 37, 49

[168] H. Zhang, Z. Duan, X. Wang , *et al.*, "Nexus-gen: A unified model for image understanding, generation, and editing," *arXiv preprint arXiv:2504.21356*, 2025. 11, 15, 24, 29, 30, 37

[169] I. AI, B. Gong, C. Zou , *et al.*, "Ming-lite-uni: Advancements in unified architecture for natural multimodal interaction," *arXiv preprint arXiv:2505.02471*, 2025. 11, 16, 30, 37

[170] H. Lin, T. Wang, Y. Ge , *et al.*, "Toklip: Marry visual tokens to clip for multimodal comprehension and generation," *arXiv preprint arXiv:2505.05422*, 2025. 11, 18, 19, 24, 28, 30, 34, 37, 67

[171] R. An, S. Yang, R. Zhang , *et al.*, "Unictokens: Boosting personalized understanding and generation via unified concept tokens," *arXiv preprint arXiv:2505.14671*, 2025. 11, 18, 21, 34, 37, 44, 49

[172] L. Yang, Y. Tian, B. Li , *et al.*, "Mmada: Multimodal large diffusion language models," *arXiv preprint arXiv:2505.15809*, 2025. 11, 18, 20, 34, 36, 37, 39, 45, 65, 66

[173] S. Wu, Z. Wu, Z. Gong , *et al.*, "Openuni: A simple baseline for unified multimodal understanding and generation," *arXiv preprint arXiv:2505.23661*, 2025. 11, 16, 30, 31, 37

[174] I. AI, B. Gong, C. Zou , *et al.*, "Ming-omni: A unified multimodal model for perception and generation," *arXiv preprint arXiv:2506.09344*, 2025. 11, 16, 31, 32, 36, 37

[175] J. Xie, Z. Yang, and M. Z. Shou, "Show-o2: Improved native unified multimodal models," *arXiv preprint arXiv:2506.15564*, 2025. 11, 19, 24, 29, 32, 33, 37, 45, 67, 68

[176] C. Wu, P. Zheng, R. Yan , *et al.*, "Omnigen2: Exploration to advanced multimodal generation," *arXiv preprint arXiv:2506.18871*, 2025. 11, 14, 29, 30, 32, 37

[177] G.-H. Wang, S. Zhao, X. Zhang , *et al.*, "Ovis-u1 technical report," *arXiv preprint arXiv:2506.23044*, 2025. 11, 29, 30, 32, 37

[178] J. Chen, Z. Cai, P. Chen , *et al.*, "Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation," *arXiv preprint arXiv:2506.18095*, 2025. 11, 18, 37, 49, 53, 68

[179] Z. Geng, Y. Wang, Y. Ma , *et al.*, "X-omni: Reinforcement learning makes discrete autoregressive image generative models great again," *arXiv preprint arXiv:2507.22058*, 2025. 11, 37, 50

[180] Z. Tan, H. Yang, L. Qin , *et al.*, "Omni-video: Democratizing unified video understanding and generation," *arXiv preprint arXiv:2507.06119*, 2025. 11

[181] P. Wang, Y. Peng, Y. Gan , *et al.*, "Skywork unipic: Unified autoregressive modeling for visual understanding and generation," *arXiv preprint arXiv:2508.03320*, 2025. 11, 37, 50

[182] S. M. Team, "Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model," 2025. [Online]. Available: https://github.com/SkyworkAI/UniPic/blob/main/UniPic-2/assets/pdf/UNIPIC2.pdf 11, 50, 68

[183] X. Wang, B. Zhuang, and Q. Wu, "Switchgpt: Adapting large language models for non-text outputs," *arXiv preprint arXiv:2309.07623*, 2023. 10, 12, 37, 42

[184] Q. Sun, Y. Wang, C. Xu , *et al.*, "Multimodal dialogue response generation," in *ACL (Volume 1: Long Papers)*, 2022. 13, 26, 34, 37, 43

[185] F. Kong, P. Wang, S. Feng , *et al.*, "Tiger: A unified generative model framework for multimodal dialogue response generation," in *LREC-COLING*, 2024. 13, 37

[186] X. Zhao, B. Liu, Q. Liu , *et al.*, "Easygen: Easing multimodal generation with bidiffuser and llms," *arXiv preprint arXiv:2310.08949*, 2023. 13, 30, 31, 37, 38, 41

[187] Z. Wang, L. Wang, Z. Zhao , *et al.*, "Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation," in *ACM MM*, 2024. 13, 36, 37, 39

[188] X. Wang, B. Zhuang, and Q. Wu, "Modaverse: Efficiently transforming modalities with llms," in *CVPR*, 2024. 13, 37, 41

[189] B. Zhu, M. Ning, P. Jin , *et al.*, "Llmbind: A unified modality-task integration framework," *arXiv preprint arXiv:2402.14891*, 2024. 13, 37, 38, 40, 42, 44, 49, 65

[190] J. Lai, J. Zhang, J. Liu , *et al.*, "Spider: Any-to-many multimodal llm," *arXiv preprint arXiv:2411.09439*, 2024. 13, 30, 31, 33

[191] H. Fei, S. Wu, H. Zhang , *et al.*, "Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing," *NeurIPS*, 2024. 13, 30, 31, 32, 37, 65

[192] Q. Guo, K. Song, Z. Feng , *et al.*, "M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance," *arXiv preprint arXiv:2502.18778*, 2025. 13, 16, 26, 36, 37, 47, 49

[193] Z. Du, Q. Chen, S. Zhang , *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024. 14, 30, 36

[194] J. Y. Koh, D. Fried, and R. R. Salakhutdinov, "Generating images with multimodal language models," *NeurIPS*, 2023. 14, 30, 31, 36, 37, 38, 40, 41, 44

[195] J. Zhu, X. Ding, Y. Ge , *et al.*, "Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation," *arXiv preprint arXiv:2312.09251*, 2023. 14, 31, 37, 48

[196] J. Li, D. Li, S. Savarese , *et al.*, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023. 14, 24, 25, 35, 40

[197] H. Tang, C. Xie, X. Bao , *et al.*, "UniLiP: Adapting CLIP for Unified Multimodal Understanding, Generation and Editing," 2025. [Online]. Available: http://arxiv.org/abs/2507.23278 14

[198] J. Wu, M. Zhong, S. Xing , *et al.*, "Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks," *NeurIPS*, 2024. 15, 31, 32, 37, 41

[199] Z. Huang, S. Zhuang, C. Fu , *et al.*, "Wegen: A unified model for interactive multimodal generation as we chat," in *CVPR*, 2025. 15, 30, 31, 37

[200] B. Lin, Z. Li, X. Cheng , *et al.*, "Uniworld: High-resolution semantic encoders for unified visual understanding and generation," *arXiv preprint arXiv:2506.03147*, 2025. 15, 24, 30, 31, 32, 65

[201] Z. Xu, J. Chen, Z. Lin , *et al.*, "Pisces: An auto-regressive foundation model for image understanding and generation," *arXiv preprint arXiv:2506.10395*, 2025. 15, 24, 30, 31, 37

[202] Z. Wang, K. Zhu, C. Xu , *et al.*, "Mio: A foundation model on multimodal tokens," *arXiv preprint arXiv:2409.17692*, 2024. 15, 27, 35, 37, 40, 42, 68

[203] R. Huang, C. Wang, J. Yang , *et al.*, "Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement," *arXiv preprint arXiv:2504.01934*, 2025. 15, 24, 29, 30, 35, 37

[204] K. Pan, W. Lin, Z. Yue , *et al.*, "Generative multimodal pretraining with discrete diffusion timestep tokens," in *CVPR*, 2025. 16, 18, 19, 24, 30, 35, 37, 44, 67

[205] H. Ye, D.-A. Huang, Y. Lu , *et al.*, "X-vila: Cross-modality alignment for large language model," *arXiv preprint arXiv:2405.19335*, 2024. 16, 29, 30, 32, 37, 41

[206] Z. Wang, Q. Duan, Y.-W. Tai , *et al.*, "C3llm: Conditional multimodal content generation using large language models," *arXiv preprint arXiv:2405.16136*, 2024. 16, 27, 35, 36, 37, 38

[207] S. Mehta, N. Jojic, and H. Gamper, "Make some noise: Towards llm audio reasoning and generation using sound tokens," in *ICASSP*, 2025. 16, 24, 30, 36, 37

[208] H. Tang, H. Liu, and X. Xiao, "Ugen: Unified autoregressive multimodal model with progressive vocabulary learning," *arXiv preprint arXiv:2503.21193*, 2025. 17, 18, 34, 37

[209] J. Sun, Y. Feng, C. Li , *et al.*, "Armor v0. 1: Empowering autoregressive multimodal understanding model with interleaved

multimodal generation via asymmetric synergy," *arXiv preprint arXiv:2503.06542*, 2025. 17, 18, 37

[210] Y. Cui, H. Chen, H. Deng , *et al.*, "Emu3.5: Native multimodal models are world learners," *arXiv preprint arXiv:2510.26583*, 2025. 17, 18, 37, 50

[211] L. Fan, L. Tang, S. Qin , *et al.*, "Unified autoregressive visual generation and understanding with continuous tokens," *arXiv preprint arXiv:2503.13436*, 2025. 17, 18, 19, 24, 25, 33, 37, 41

[212] A. Aghajanyan, B. Huang, C. Ross , *et al.*, "Cm3: A causal masked multimodal model of the internet," *arXiv preprint arXiv:2201.07520*, 2022. 18, 34, 37, 40, 44, 48

[213] M. Yasunaga, A. Aghajanyan, W. Shi , *et al.*, "Retrieval-augmented multimodal language modeling," in *ICML*, 2023. 18, 26, 34, 37, 47

[214] L. Yu, B. Shi, R. Pasunuru , *et al.*, "Scaling autoregressive multimodal models: Pretraining and instruction tuning," *arXiv preprint arXiv:2309.02591*, 2023. 18, 34, 37, 44

[215] R. Xie, C. Du, P. Song , *et al.*, "Muse-vl: Modeling unified vlm through semantic discrete encoding," in *ICCV*, 2025. 18, 24, 28, 30, 34, 37, 67

[216] Z. Chen, C. Wang, X. Chen , *et al.*, "Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation," *arXiv preprint arXiv:2503.06764*, 2025. 18, 19, 24, 28, 30, 34, 37, 67

[217] X. V. Lin, A. Shrivastava, L. Luo , *et al.*, "Moma: Efficient early-fusion pre-training with mixture of modality-aware experts," *arXiv preprint arXiv:2407.21770*, 2024. 18, 19, 22, 37, 66

[218] Z. Sun, Z. Chu, P. Zhang , *et al.*, "X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models," *arXiv preprint arXiv:2412.01824*, 2024. 18, 19, 37, 65

[219] J. Yang, D. Yin, Y. Zhou , *et al.*, "Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling," in *CVPR*, 2025. 18, 19, 33, 37, 66

[220] S. Kou, J. Jin, Z. Liu , *et al.*, "Orthus: Autoregressive interleaved image-text generation with modality-specific heads," *arXiv preprint arXiv:2412.00127*, 2024. 18, 19, 33, 37, 66

[221] Y. Sun, H. Bao, W. Wang , *et al.*, "Multimodal latent language modeling with next-token diffusion," *arXiv preprint arXiv:2412.08635*, 2024. 18, 19, 24, 30, 33, 37, 40

[222] P. Wang, Y. Peng, Y. Gan , *et al.*, "Skywork unipic: Unified autoregressive modeling for visual understanding and generation," *arXiv preprint arXiv:2508.03320*, 2025. 18, 19, 66

[223] J. Zhang, Y. Liu, Y.-W. Tai , *et al.*, "C3net: Compound conditioned controlnet for multimodal content generation," in *CVPR*, 2024. 18, 20, 37, 42

[224] M. Hu, C. Zheng, Z. Yang , *et al.*, "Unified discrete diffusion for simultaneous vision-language generation," in *ICLR*, 2023. 18, 20

[225] Z. Li, H. Li, Y. Shi , *et al.*, "Dual diffusion for unified image generation and understanding," in *CVPR*, 2025. 18, 20, 33, 37, 47, 65

[226] C. Xu, X. Wang, Z. Liao , *et al.*, "Unicms: A unified consistency model for efficient multimodal generation and understanding," *arXiv preprint arXiv:2502.05415*, 2025. 18, 20, 22, 34, 37, 43, 45, 65

[227] S. Li, K. Kallidromitis, A. Gokul , *et al.*, "Omniflow: Any-to-any generation with multi-modal rectified flows," in *CVPR*, 2025. 18, 20, 24, 33, 37, 47

[228] C. Zhao, Y. Song, W. Wang , *et al.*, "Monoformer: One transformer for both diffusion and autoregression," *arXiv preprint arXiv:2409.16280*, 2024. 18, 19, 21, 32, 37

[229] R. Zhao, W. Mao, and M. Z. Shou, "Doracycle: Domain-oriented adaptation of unified generative model in multimodal cycles," in *CVPR*, 2025. 18, 21, 34, 37, 40, 42

[230] R. Tian, M. Gao, M. Xu , *et al.*, "Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation," *arXiv preprint arXiv:2505.14682*, 2025. 18, 21, 34, 37

[231] W. Shi, X. Han, C. Zhou , *et al.*, "Llamafusion: Adapting pretrained language models for multimodal generation," *arXiv preprint arXiv:2412.15188*, 2024. 18, 22, 32, 37, 42, 66

[232] S. Mo, T. Nguyen, X. Huang , *et al.*, "X-fusion: Introducing new modality to frozen large language models," *arXiv preprint arXiv:2504.20996*, 2025. 18, 22, 32, 37

[233] Y. Huang, H. Xue, B. Liu , *et al.*, "Unifying multimodal transformer for bi-directional image and text generation," in *ACM MM*, 2021. 18, 23

[234] S. Diao, W. Zhou, X. Zhang , *et al.*, "Write and paint: Generative vision-language models are unified modal learners," *arXiv preprint arXiv:2206.07699*, 2022. 18, 23, 34, 37

[235] Y. Fang, B. Jin, J. Shen , *et al.*, "Graphgpt-o: Synergistic multimodal comprehension and generation on graphs," in *CVPR*, 2025. 18, 23, 37

[236] A. Dubey, A. Jauhri, A. Pandey , *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024. 17, 22

[237] G. Team, M. Riviere, S. Pathak , *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024. 17

[238] Qwen, :, A. Yang , *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2025. 17

[239] P. Sun, Y. Jiang, S. Chen , *et al.*, "Autoregressive model beats diffusion: Llama for scalable image generation," *arXiv preprint arXiv:2406.06525*, 2024. 18, 23, 34

[240] T. Li, Y. Tian, H. Li , *et al.*, "Autoregressive image generation without vector quantization," *NeurIPS*, 2024. 19, 33, 64

[241] J. Austin, D. D. Johnson, J. Ho , *et al.*, "Structured denoising diffusion models in discrete state-spaces," *NeurIPS*, 2021. 20

[242] A. Lou, C. Meng, and S. Ermon, "Discrete diffusion modeling by estimating the ratios of the data distribution," *arXiv preprint arXiv:2310.16834*, 2023. 20

[243] S. Nie, F. Zhu, Z. You , *et al.*, "Large language diffusion models," *arXiv preprint arXiv:2502.09992*, 2025. 21

[244] F. Zhu, R. Wang, S. Nie , *et al.*, "Llada 1.5: Variance-reduced preference optimization for large language diffusion models," *arXiv preprint arXiv:2505.19223*, 2025. 21

[245] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022. 21, 32, 33

[246] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023. 22

[247] A. Vaswani, N. Shazeer, N. Parmar , *et al.*, "Attention is all you need," *NeurIPS*, 2017. 23

[248] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024. 23

[249] M. Oquab, T. Darcet, T. Moutakanni , *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," 2024. [Online]. Available: http://arxiv.org/abs/2304.07193 23

[250] O. Rybkin, K. Daniilidis, and S. Levine, "Simple and effective vae training with calibrated decoders," *arXiv preprint arXiv:2006.13202*, 2021. 23, 24

[251] M. Dehghani, B. Mustafa, J. Djolonga , *et al.*, "Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution," *arXiv preprint arXiv:2307.06304*, 2023. 24

[252] Y. Zhu, Y. Zhou, C. Wang , *et al.*, "Unit: Unifying image and text recognition in one vision encoder," *arXiv preprint arXiv:2409.04095*, 2024. 24

[253] Z. Liu, J. Ning, Y. Cao , *et al.*, "Video swin transformer," in *CVPR*, 2022. 24, 25, 38

[254] A. Arnab, M. Dehghani, G. Heigold , *et al.*, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021. 24, 25

[255] Z. Gao, S. Zhang, M. Lei , *et al.*, "San-m: Memory equipped self-attention for end-to-end speech recognition," *arXiv preprint arXiv:2006.01713*, 2020. 24, 26

[256] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *Interspeech 2021*, 2021. 24, 26, 38

[257] Y. Wu*, K. Chen*, T. Zhang* , *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023. 24, 32

[258] Y. Li, R. Yuan, G. Zhang , *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2024. 24, 26

[259] D. Lee, C. Kim, S. Kim , *et al.*, "Autoregressive image generation using residual quantization," in *CVPR*, 2022. 24, 26, 30, 38

[260] C. Zheng, T.-L. Vuong, J. Cai , *et al.*, "Movq: Modulating quantized vectors for high-fidelity image generation," *NeurIPS*, 2022. 24, 26, 27, 30, 35

[261] V. Iashin and E. Rahtu, "Taming visually guided sound generation," *arXiv preprint arXiv:2110.08791*, 2021. 24

[262] A. Défossez, J. Copet, G. Synnaeve , *et al.*, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022. 24, 27, 30, 35

[263] Z. Xin, Z. Dong, L. Shimin , *et al.*, "Speechtokenizer: Unified speech tokenizer for speech language models," in *Proc. Int. Conf. Learn. Representations*, 2024. 24, 27, 30, 35

[264] H. Li, C. Tian, J. Shao , *et al.*, "Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding," in *CVPR*, 2025. 24, 30, 34, 37, 42

[265] F. Yang, Y. Zhu, X. Li , *et al.*, "Focus: Unified vision-language modeling for interactive editing driven by referential segmentation," *arXiv preprint arXiv:2506.16806*, 2025. 24, 29, 30, 35, 37

[266] J. Han, H. Chen, Y. Zhao , *et al.*, "Vision as a dialect: Unifying visual understanding and generation via text-aligned representations," *arXiv preprint arXiv:2506.18898*, 2025. 24, 28, 30, 35, 37

[267] J. Liu, S. Chen, X. He , *et al.*, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *IEEE TPAMI*, 2024. 26, 37, 38, 41, 42

[268] Y. Chen, H. Zhong, Y. Li , *et al.*, "Unicode^2: Cascaded large-scale codebooks for unified multimodal understanding and generation," *arXiv preprint arXiv:2506.20214*, 2025. 26, 28, 35, 37

[269] M. Tschannen, A. Gritsenko, X. Wang , *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025. 28

[270] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023. 30, 32, 51, 55, 56

[271] Y. Li, H. Liu, Q. Wu , *et al.*, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023. 30

[272] E. Xie, J. Chen, J. Chen , *et al.*, "Sana: Efficient high-resolution image synthesis with linear diffusion transformers," *arXiv preprint arXiv:2410.10629*, 2024. 30, 31

[273] Q. Qin, L. Zhuo, Y. Xin , *et al.*, "Lumina-image 2.0: A unified and efficient image generative framework," *arXiv preprint arXiv:2503.21758*, 2025. 30

[274] cerspense, "zeroscope," 2023. [Online]. Available: https://huggingface.co/cerspense 30, 32

[275] S. Zhang, J. Wang, Y. Zhang , *et al.*, "I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models," *arXiv preprint arXiv:2311.04145*, 2023. 30, 32

[276] H. Chen, Y. Zhang, X. Cun , *et al.*, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," *arXiv preprint arXiv:2401.09047*, 2024. 30, 32

[277] T. Wan, A. Wang, B. Ai , *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025. 30, 32, 33, 67

[278] H. Liu, Z. Chen, Y. Yuan , *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023. 30, 32, 33

[279] H. Liu, Y. Yuan, X. Liu , *et al.*, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *TASLP*, 2024. 30, 32

[280] O. Gafni, A. Polyak, O. Ashual , *et al.*, "Make-a-scene: Scene-based text-to-image generation with human priors," in *ECCV*, 2022. 30, 34

[281] J. Han, J. Liu, Y. Jiang , *et al.*, "Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis," *arXiv preprint arXiv:2412.04431*, 2024. 30

[282] L. Yu, J. Lezama, N. B. Gundavarapu , *et al.*, "Language model beats diffusion – tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2024. 30, 34, 39, 67

[283] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. 31, 34

[284] L. Zhuo, R. Du, H. Xiao , *et al.*, "Lumina-next: Making lumina-t2x stronger and faster with next-dit," *arXiv preprint arXiv:2406.18583*, 2024. 31

[285] X. Zhu, W. Su, L. Lu , *et al.*, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020. 32

[286] T. Bai, J. Luo, J. Zhao , *et al.*, "Recent advances in adversarial training for adversarial robustness," *arXiv preprint arXiv:2102.01356*, 2021. 32

[287] H. Chen, M. Xia, Y. He , *et al.*, "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023. 32

[288] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *NeurIPS*, 2020. 32, 33

[289] P.-Y. Huang, H. Xu, J. Li , *et al.*, "Masked autoencoders that listen," *NeurIPS*, 2022. 32

[290] Z. Liu, H. Mao, C.-Y. Wu , *et al.*, "A convnet for the 2020s," *CVPR*, 2022. 33

[291] W. Shi, J. Caballero, F. Huszár , *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016. 33

[292] L. Fan, T. Li, S. Qin , *et al.*, "Fluid: Scaling autoregressive text-to-image generative models with continuous tokens," *arXiv preprint arXiv:2410.13863*, 2024. 33

[293] B. Chen, D. Martí Monsó, Y. Du , *et al.*, "Diffusion forcing: Next-token prediction meets full-sequence diffusion," *NeurIPS*, 2024. 33

[294] J. Ao, R. Wang, L. Zhou , *et al.*, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *ACL (Volume 1: Long Papers)*, 2022. 33

[295] S. Patil, W. Berman, R. Rombach , *et al.*, "amused: An open muse reproduction," *arXiv preprint arXiv:2401.01808*, 2024. 34

[296] A. Forever, "Sber-movqgan," 2023. [Online]. Available: https://github.com/ai-forever/MoVQGAN 34, 40

[297] J. Yu, X. Li, J. Y. Koh , *et al.*, "Vector-quantized image modeling with improved vqgan," *arXiv preprint arXiv:2110.04627*, 2021. 35

[298] D. Le Gall, "Mpeg: A video compression standard for multimedia applications," *Communications of the ACM*, 1991. 35

[299] J. Y. Koh, R. Salakhutdinov, and D. Fried, "Grounding language models to images for multimodal inputs and outputs," in *ICML*, 2023. 36, 37, 38, 41, 44, 48

[300] S. Lee, W. J. Kim, J. Chang , *et al.*, "Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation," in *ICLR*, 2024. 37, 38, 48, 65

[301] W. Chow, J. Li, Q. Yu , *et al.*, "Unified generative and discriminative training for multi-modal large language models," *NeurIPS*, 2024. 37, 40, 42, 44

[302] Z. Wang, J. Lorraine, Y. Wang , *et al.*, "Llama-mesh: Unifying 3d mesh generation with language models," *arXiv preprint arXiv:2411.09595*, 2024. 37, 65

[303] L. Yang, Z. Wang, Z. Chen , *et al.*, "Medxchat: A unified multimodal large language model framework towards cxrs understanding and generation," in *ISBI*, 2025. 37, 39, 41, 65

[304] T. Lin, W. Zhang, S. Li , *et al.*, "Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation," *ICML*, 2025. 37, 65

[305] W. Mao, Z. Yang, and M. Z. Shou, "Unimod: Efficient unified multimodal transformers with mixture-of-depths," *arXiv preprint arXiv:2502.06474*, 2025. 37, 44

[306] Z. Zhou, Y. Zhu, M. Zhu , *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," *arXiv preprint arXiv:2502.14420*, 2025. 37, 64

[307] T. Li, Q. Lu, L. Zhao , *et al.*, "Unifork: Exploring modality alignment for unified multimodal understanding and generation," *arXiv preprint arXiv:2506.17202*, 2025. 37, 49

[308] J. Jiang, C. Si, J. Luo , *et al.*, "Co-reinforcement learning for unified multimodal understanding and generation," *arXiv preprint arXiv:2505.17534*, 2025. 37

[309] Y. Xin, Q. Qin, S. Luo , *et al.*, "Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding," *arXiv preprint arXiv:2510.06308*, 2025. 37

[310] OpenAI, "Dall·e 3 is now available in chatgpt plus and enterprise," 2023. [Online]. Available: https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise 38

[311] J. Deng, W. Dong, R. Socher , *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009. 39, 40, 54, 55

[312] T.-Y. Lin, M. Maire, S. Belongie , *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, 2014. 39, 40, 51, 53, 54, 55, 59

[313] S. Y. Gadre, G. Ilharco, A. Fang , *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *NeurIPS*, 2023. 39, 55

[314] P. Sharma, N. Ding, S. Goodman , *et al.*, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL (Volume 1: Long Papers)*, 2018. 40, 50, 55, 56

[315] R. A. Jacobs, M. I. Jordan, S. J. Nowlan , *et al.*, "Adaptive mixtures of local experts," *Neural computation*, 1991. 42

[316] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *arXiv preprint arXiv:1312.4314*, 2013. 42

[317] C. Raffel, N. Shazeer, A. Roberts , *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, 2020. 43

[318] Y. Ge, S. Zhao, C. Li , *et al.*, "Seed-data-edit technical report: A hybrid dataset for instructional image editing," *arXiv preprint arXiv:2405.04007*, 2024. 46, 48, 51, 53, 55

[319] Y. Li, S. Jiang, B. Hu , *et al.*, "Uni-moe: Scaling unified multimodal llms with mixture of experts," *IEEE TPAMI*, 2025. 47

[320] E. J. Hu, Y. Shen, P. Wallis , *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, 2022. 47

[321] R. Rafailov, A. Sharma, E. Mitchell , *et al.*, "Direct preference optimization: Your language model is secretly a reward model," *NeurIPS*, 2023. 49, 50

[322] D. Guo, D. Yang, H. Zhang , *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025. 49, 68

[323] J. Hong, Y. Zhang, G. Wang , *et al.*, "Reinforcing multimodal understanding and generation with dual self-rewards," *arXiv preprint arXiv:2506.07963*, 2025. 50

[324] T. Moreau and J. Audiffren, "Post training in deep learning with last kernel," *arXiv preprint arXiv:1611.04499*, 2016. 50

[325] K. Kumar, T. Ashraf, O. Thawakar , *et al.*, "Llm post-training: A deep dive into reasoning large language models," *arXiv preprint arXiv:2502.21321*, 2025. 50

[326] G. Tie, Z. Zhao, D. Song , *et al.*, "Large language models post-training: Surveying techniques from alignment to reasoning," *arXiv preprint arXiv:2503.06072*, 2025. 50

[327] C. Schuhmann, R. Beaumont, R. Vencu , *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS*, 2022. 50, 51, 52, 54, 55, 56, 59

[328] S. Changpinyo, P. Sharma, N. Ding , *et al.*, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021. 50, 55, 56

[329] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *NeurIPS*, 2011. 51, 55

[330] B. A. Plummer, L. Wang, C. M. Cervantes , *et al.*, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015. 51

[331] M. Byeon, B. Park, H. Kim , *et al.*, "Coyo-700m: Image-text pair dataset," 2022. [Online]. Available: https://github.com/kakaobrain/coyo-dataset 51, 52, 55

[332] LAION-AI, "aesthetic-predictor," 2022. [Online]. Available: https://github.com/LAION-AI/aesthetic-predictor 51, 52

[333] C. Schuhmann , *et al.*, "Laion-aesthetics," 2022. [Online]. Available: https://laion.ai/blog/laion-aesthetics/ 51, 54, 55, 56

[334] J. Hessel, A. Holtzman, M. Forbes , *et al.*, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021. 51, 52, 56

[335] LAION-AI, "Clip-based-nsfw-detector," 2022. [Online]. Available: https://github.com/LAION-AI/CLIP-based-NSFW-Detector 51, 52

[336] C. Wu, J. Li, J. Zhou , *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025. 52

[337] J.-B. Alayrac, J. Donahue, P. Luc , *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, 2022. 52

[338] Y. Goyal, T. Khot, D. Summers-Stay , *et al.*, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *CVPR*, 2017. 53, 54, 55, 57, 58

[339] S. Kazemzadeh, V. Ordonez, M. Matten , *et al.*, "ReferItGame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014. 53, 54, 55

[340] J. Xu, X. Liu, Y. Wu , *et al.*, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *NeurIPS*, 2023. 54

[341] X. Wu, Y. Hao, K. Sun , *et al.*, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023. 54

[342] K. Zhang, L. Mo, W. Chen , *et al.*, "Magicbrush: A manually annotated dataset for instruction-guided image editing," *NeurIPS*, 2023. 55, 56

[343] C. Foundation, "Midjourney v6 dataset by bittensor network (netuid 19)," 2024. [Online]. Available: https://huggingface.co/datasets/CortexLM/midjourney-v6 55

[344] K. Nan, R. Xie, P. Zhou , *et al.*, "Openvid-1m: A large-scale high-quality dataset for text-to-video generation," *arXiv preprint arXiv:2407.02371*, 2024. 55

[345] J. Hate, "text-to-image-2m (revision e64fca4)," 2024. [Online]. Available: https://huggingface.co/datasets/jackyhate/text-to-image-2M 55

[346] S. Zhang, X. Yang, Y. Feng , *et al.*, "Hive: Harnessing human feedback for instructional visual editing," in *CVPR*, 2024. 55

[347] M. Hui, S. Yang, B. Zhao , *et al.*, "Hq-edit: A high-quality dataset for instruction-based image editing," *arXiv preprint arXiv:2404.09990*, 2024. 55

[348] Z. Chen, X. Bai, Y. Shi , *et al.*, "Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing," *arXiv preprint arXiv:2509.24900*, 2025. 55, 56

[349] H. Zhao, X. Ma, L. Chen , *et al.*, "Ultraedit: Instruction-based fine-grained image editing at scale," *arXiv preprint arXiv:2407.05282*, 2024. 55, 58, 62

[350] Y. Ye, X. He, Z. Li , *et al.*, "Imgedit: A unified image editing dataset and benchmark," *arXiv preprint arXiv:2505.20275*, 2025. 55, 58

[351] N. Kumari, X. Yin, J.-Y. Zhu , *et al.*, "Generating multi-image synthetic data for text-to-image customization," *arXiv preprint arXiv:2502.01720*, 2025. 55

[352] C. Qin, S. Zhang, N. Yu , *et al.*, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," *arXiv preprint arXiv:2305.11147*, 2023. 55

[353] L. Li, Y. Yin, S. Li , *et al.*, "M^3 it: A large-scale dataset towards multi-modal multilingual instruction tuning," *arXiv preprint arXiv:2306.04387*, 2023. 55, 56

[354] B. Li, Y. Zhang, L. Chen , *et al.*, "Mimic-it: multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023. 55, 56

[355] F. Liu, K. Lin, L. Li , *et al.*, "Mitigating hallucination in large multi-modal models via robust instruction tuning," in *ICLR*, 2024. 55

[356] B. Li, Y. Zhang, D. Guo , *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024. 55

[357] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh , *et al.*, "Visual storytelling," in *NAACL*, 2016. 55, 56, 59

[358] A. Das, S. Kottur, K. Gupta , *et al.*, "Visual Dialog," in *CVPR (CVPR)*, 2017. 55, 56

[359] Y. Zhang, R. Zhang, J. Gu , *et al.*, "Llavar: Enhanced visual instruction tuning for text-rich image understanding," *arXiv preprint arXiv:2306.17107*, 2023. 55

[360] J. Wang, L. Meng, Z. Weng , *et al.*, "To see is to believe: Prompting gpt-4v for better visual instruction tuning," *arXiv preprint arXiv:2311.07574*, 2023. 55

[361] lmms lab, "Llava-recap-cc3m," 2024. [Online]. Available: https://huggingface.co/datasets/lmms-lab/LLaVA-ReCap-CC3M 55

[362] Y. Huang, X. Sheng, Z. Yang , *et al.*, "Aesexpert: Towards multi-modality foundation model for image aesthetics perception," in *ACM MM*, 2024. 55

[363] S. Gu, J. Zhang, S. Zhou , *et al.*, "Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data," *arXiv preprint arXiv:2410.18558*, 2024. 55

[364] G. H. Chen, S. Chen, R. Zhang , *et al.*, "Allava: Harnessing gpt4v-synthesized data for a lite vision-language model," *arXiv preprint arXiv:2402.11684*, 2024. 55

[365] M. Deitke, E. VanderBilt, A. Herrasti , *et al.*, "Procthor: Large-scale embodied ai using procedural generation," *NeurIPS*, 2022. 55, 56

[366] R. Ramrakhya, E. Undersander, D. Batra , *et al.*, "Habitat-web: Learning embodied object-search strategies from human demonstrations at scale," in *CVPR*, 2022. 55, 56

[367] Y. Jiang, A. Gupta, Z. Zhang , *et al.*, "Vima: General robot manipulation with multimodal prompts," *arXiv preprint arXiv:2210.03094*, 2022. 55, 56

[368] C. Rawles, A. Li, D. Rodriguez , *et al.*, "Androidinthewild: A large-scale dataset for android device control," *NeurIPS*, 2023. 55

[369] W. Li, W. E. Bishop, A. Li , *et al.*, "On the effects of data scale on ui control agents," *NeurIPS*, 2024. 55

[370] P. Weinzaepfel, V. Leroy, T. Lucas , *et al.*, "Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion," *NeurIPS*, 2022. 55, 56

[371] M. Deitke, D. Schwenk, J. Salvador , *et al.*, "Objaverse: A universe of annotated 3d objects," in *CVPR*, 2023. 55, 56

[372] Y. Yao, Z. Luo, S. Li , *et al.*, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," in *CVPR*, 2020. 55, 56

[373] X. Han, Y. Wu, L. Shi , *et al.*, "Mvimgnet2. 0: A larger-scale dataset of multi-view images," *arXiv preprint arXiv:2412.01430*, 2024. 55

[374] H. Laurençon, L. Saulnier, L. Tronchon , *et al.*, "Obelics: An open web-scale filtered dataset of interleaved image-text documents," *NeurIPS*, 2023. 55, 56

[375] W. Zhu, J. Hessel, A. Awadalla , *et al.*, "Multimodal C4: An open, billion-scale corpus of images interleaved with text," *arXiv preprint arXiv:2304.06939*, 2023. 55, 56

[376] M. Koupaee and W. Y. Wang, "Wikihow: A large scale text summarization dataset," *arXiv preprint arXiv:1810.09305*, 2018. 55, 56, 59

[377] K. Srinivasan, K. Raman, J. Chen , *et al.*, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *ACM SIGIR*, 2021. 55

[378] Q. Li, Z. Chen, W. Wang , *et al.*, "Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text," in *ICLR*, 2025. 55

[379] A. Hu, H. Xu, J. Ye , *et al.*, "mplug-docowl 1.5: Unified structure learning for ocr-free document understanding," *arXiv preprint arXiv:2403.12895*, 2024. 55

[380] V. Singla, K. Yue, S. Paul , *et al.*, "From pixels to prose: A large dataset of dense image captions," *arXiv preprint arXiv:2406.10328*, 2024. 55

[381] W. Chen, L. Li, Y. Yang , *et al.*, "Comm: A coherent interleaved image-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2406.10462*, 2024. 55, 58, 59, 63

[382] H. Chen, W. Xie, A. Vedaldi , *et al.*, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020. 55, 56

[383] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017. 55, 56

[384] J. F. Gemmeke, D. P. Ellis, D. Freedman , *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017. 55, 56

[385] K. Ito and L. Johnson, "The lj speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/ 55, 56

[386] V. Pratap, Q. Xu, A. Sriram , *et al.*, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020. 55, 56

[387] G. Chen, S. Chai, G. Wang , *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021. 55

[388] F. Font, A. Mesaros, D. P. Ellis , *et al.*, "Proceedings of the 6th workshop on detection and classication of acoustic scenes and events," in *DCASE Workshop*, 2021. 55

[389] C. D. Kim, B. Kim, H. Lee , *et al.*, "Audiocaps: Generating captions for audios in the wild," in *NAACL on Human Language Technologies (Long and Short Papers)*, 2019. 55, 56

[390] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP*, 2020. 55, 56

[391] A. Agostinelli, T. I. Denk, Z. Borsos , *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023. 55, 56

[392] W. Kang, X. Yang, Z. Yao , *et al.*, "Libriheavy: A 50,000 hours asr corpus with punctuation casing and context," in *ICASSP*, 2024. 55

[393] ogbanugot, "musicgen," 2024. [Online]. Available: https://huggingface.co/datasets/ogbanugot/musicgen 55

[394] X. Wang, J. Wu, J. Chen , *et al.*, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *ICCV*, 2019. 55

[395] M. Bain, A. Nagrani, G. Varol , *et al.*, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021. 55, 59

[396] J. Xu, T. Mei, T. Yao , *et al.*, "Msr-vtt: A large video description dataset for bridging video and language," in *CVPR*, 2016. 55, 59

[397] E. Real, J. Shlens, S. Mazzocchi , *et al.*, "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video," in *CVPR*, 2017. 55

[398] D. Yang, S. Huang, C. Lu , *et al.*, "Vript: A video is worth thousands of words," *NeurIPS*, 2024. 55

[399] Y. Wang, Y. He, Y. Li , *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023. 55, 56

[400] A. Miech, D. Zhukov, J.-B. Alayrac , *et al.*, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019. 55

[401] H. Fan, H. Bai, L. Lin , *et al.*, "Lasot: A high-quality large-scale single object tracking benchmark," *IJCV*, 2021. 55

[402] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE TPAMI*, 2019. 55

[403] C. Gu, C. Sun, D. Ross , *et al.*, "Ava: a video dataset of spatio-temporally localized atomic visual actions (2018)," *arXiv preprint arXiv:1705.08421*, 2017. 55

[404] Q. Wang, Y. Shi, J. Ou , *et al.*, "Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content," in *CVPR*, 2025. 55

[405] Y. Zhang, J. Wu, W. Li , *et al.*, "Llava-video: Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024. 55

[406] R. Zellers, X. Lu, J. Hessel , *et al.*, "Merlot: Multimodal neural script knowledge models," in *NeurIPS 34*, 2021. 54, 55

[407] R. Zellers, J. Lu, X. Lu , *et al.*, "Merlot reserve: Neural script knowledge through vision and language and sound," in *CVPR*, 2022. 55

[408] S. Lee, J. Chung, Y. Yu , *et al.*, "Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning," in *ICCV*, 2021. 55

[409] H. Xue, T. Hang, Y. Zeng , *et al.*, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *CVPR*, 2022. 55

[410] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012. 54, 55

[411] K. Grauman, A. Westbury, E. Byrne , *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022. 55

[412] W. Kay, J. Carreira, K. Simonyan , *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 55, 59

[413] R. Goyal, S. Ebrahimi Kahou, V. Michalski , *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *ICCV*, 2017. 55

[414] D. Damen, H. Doughty, G. M. Farinella , *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *IJCV*, 2022. 55, 58

[415] F. Hamborg, N. Meuschke, C. Breitinger , *et al.*, "news-please: A generic news crawler and extractor," in *ISI*, 2017. 54, 55

[416] Wikimedia Foundation, "Wikipedia database dump." [Online]. Available: https://dumps.wikimedia.org/ 54, 55

[417] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019. 54, 55

[418] D. Gurari, Q. Li, A. J. Stangl , *et al.*, "Vizwiz grand challenge: Answering visual questions from blind people," *arXiv preprint arXiv:1802.08218*, 2018. 54, 55, 57, 58

[419] K. Marino, M. Rastegari, A. Farhadi , *et al.*, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019. 55

[420] D. Schwenk, A. Khandelwal, C. Clark , *et al.*, "A-okvqa: A benchmark for visual question answering using world knowledge," in *ECCV*, 2022. 54, 55

[421] S. Li and N. Tajbakhsh, "Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs," *arXiv preprint arXiv:2308.03349*, 2023. 55

[422] M. Acharya, K. Kafle, and C. Kanan, "Tallyqa: Answering complex counting questions," in *AAAI*, 2019. 55

[423] A. Mishra, S. Shekhar, A. K. Singh , *et al.*, "Ocr-vqa: Visual question answering by reading text in images," in *ICDAR*, 2019. 55

[424] P. Lu, S. Mishra, T. Xia , *et al.*, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *NeurIPS*, 2022. 54, 55, 57, 58

[425] R. Zellers, Y. Bisk, A. Farhadi , *et al.*, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, 2018. 55

[426] D. Xu, Z. Zhao, J. Xiao , *et al.*, "Video question answering via gradually refined attention over appearance and motion," in *ACM MM*, 2017. 54, 55, 58

[427] J. Gao, R. Pi, J. Zhang , *et al.*, "G-llava: Solving geometric problem with multi-modal large language model," *arXiv preprint arXiv:2312.11370*, 2023. 55

[428] B. Wu, S. Yu, Z. Chen , *et al.*, "Star: A benchmark for situated reasoning in real-world videos," *arXiv preprint arXiv:2405.09711*, 2024. 55

[429] M. Zhao, B. Li, J. Wang , *et al.*, "Towards video text visual question answering: Benchmark and baseline," *NeurIPS*, 2022. 55

[430] K. Kafle, B. Price, S. Cohen , *et al.*, "Dvqa: Understanding data visualizations via question answering," in *CVPR*, 2018. 55

[431] A. Masry, D. Long, J. Q. Tan , *et al.*, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *ACL*, 2022. 55, 58, 60

[432] X. Chen, Z. Zhao, L. Chen , *et al.*, "WebSRC: A dataset for web-based structural reading comprehension," in *EMNLP*, 2021. 55, 58, 59

[433] C. Schuhmann, R. Vencu, R. Beaumont , *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021. 54, 55

[434] X. Chen, X. Wang, S. Changpinyo , *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022. 55

[435] Z. Tan, S. Liu, X. Yang , *et al.*, "Ominicontrol: Minimal and universal control for diffusion transformer," *arXiv preprint arXiv:2411.15098*, 2024. 55

[436] C. Schuhmann, A. Köpf, R. Vencu , *et al.*, "Laion-coco: 600m machine-generated captions for the coco dataset," 2022. [Online]. Available: https://laion.ai/blog/laion-coco/ 55

[437] K. Desai, G. Kaul, Z. Aysola , *et al.*, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv preprint arXiv:2111.11431*, 2021. 55, 56

[438] X. Jin, Q. Qiao, Y. Lu , *et al.*, "Apddv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments," *arXiv preprint arXiv:2411.08545*, 2024. 55

[439] X. Li, F. Zhang, H. Diao , *et al.*, "Densefusion-1m: Merging vision experts for comprehensive multimodal perception," *NeurIPS*, 2024. 55

[440] A. Kuznetsova, H. Rom, N. Alldrin , *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020. 55

[441] S. Shao, Z. Li, T. Zhang , *et al.*, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV (ICCV)*, 2019. 54, 55, 59

[442] B. Thomee, D. A. Shamma, G. Friedland , *et al.*, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, 2016. 55, 59

[443] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019. 55

[444] A. Suhr, M. Lewis, J. Yeh , *et al.*, "A corpus of natural language for visual reasoning," in *ACL (Volume 2: Short Papers)*, 2017. 55

[445] A. Suhr, S. Zhou, A. Zhang , *et al.*, "A corpus for reasoning about natural language grounded in photographs," *arXiv preprint arXiv:1811.00491*, 2018. 55

[446] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *TACL*, 2023. 55

[447] pixparse, "pdfa-eng-wds," 2024. [Online]. Available: https://huggingface.co/datasets/pixparse/pdfa-eng-wds 55

[448] G. Brazil, A. Kumar, J. Straub , *et al.*, "Omni3D: A large benchmark and model for 3D object detection in the wild," in *CVPR*, 2023. 55

[449] S. Schulter, Y. Suh, K. M. Dafnis , *et al.*, "Omnilabel: A challenging benchmark for language-based object detection," in *ICCV*, 2023. 55

[450] A. Veit, T. Matera, L. Neumann , *et al.*, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," in *arXiv preprint arXiv:1601.07140*, 2016. 55

[451] A. R. Zamir, A. Sax, W. Shen , *et al.*, "Taskonomy: Disentangling task transfer learning," in *CVPR*, 2018. 55

[452] A. Singh, G. Pang, M. Toh , *et al.*, "Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *CVPR*, 2021. 55

[453] K. Sun, J. Pan, Y. Ge , *et al.*, "Journeydb: A benchmark for generative image understanding," *NeurIPS*, 2023. 55, 56

[454] H. Caesar, J. R. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context." *arXiv preprint arXiv:1612.03716*, 2016. 55

[455] V. Ramanathan, A. Kalia, V. Petrovic , *et al.*, "Paco: Parts and attributes of common objects," in *CVPR*, 2023. 55

[456] J. He, S. Yang, S. Yang , *et al.*, "Partimagenet: A large, high-quality dataset of parts," in *ECCV*, 2022. 55

[457] X. Chen, R. Mottaghi, X. Liu , *et al.*, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *CVPR*, 2014. 55

[458] laion, "gpt4v-dataset," 2023. [Online]. Available: https://huggingface.co/datasets/jackyhate/text-to-image-2M 55

[459] B. Zhou, H. Zhao, X. Puig , *et al.*, "Scene parsing through ade20k dataset," in *CVPR*, 2017. 55

[460] C. Luke, C. Timothy, and Z. Ali, "Unsplash lite dataset." [Online]. Available: https://unsplash.com/data 55

[461] ——, "Unsplash full dataset." [Online]. Available: https://unsplash.com/data 55, 56

[462] J. Chen, Y. Huang, T. Lv , *et al.*, "Textdiffuser: Diffusion models as text painters," *NeurIPS*, 2023. 55

[463] L. Chen, J. Li, X. Dong , *et al.*, "Sharegpt4v: Improving large multi-modal models with better captions," *arXiv preprint arXiv:2311.12793*, 2023. 54, 55

[464] Q. Yu, Q. Sun, X. Zhang , *et al.*, "Capsfusion: Rethinking image-text data at scale," in *CVPR*, 2024. 55

[465] Z. Peng, W. Wang, L. Dong , *et al.*, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023. 55

[466] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012. 54, 55

[467] O. Russakovsky, J. Deng, H. Su , *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015. 54, 55, 59

[468] B. Zhou, A. Lapedriza, A. Khosla , *et al.*, "Places: A 10 million image database for scene recognition," *IEEE TPAMI*, 2017. 54, 55

[469] J. Xiao, J. Hays, K. A. Ehinger , *et al.*, "SUN database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010. 55

[470] Y. Cui, Y. Song, C. Sun , *et al.*, "Large scale fine-grained categorization and domain-specific transfer learning," in *CVPR*, 2018. 55

[471] A. Dosovitskiy, P. Fischer, E. Ilg , *et al.*, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015. 54, 55

[472] C. Wah, S. Branson, P. Welinder , *et al.*, "The caltech-ucsd birds-200-2011 dataset," california institute of technology, Tech. Rep., 2011. 55

[473] A. Kirillov, E. Mintun, N. Ravi , *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 55

[474] D. J. Butler, J. Wulff, G. B. Stanley , *et al.*, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012. 54, 55

[475] S. He, Y. Zhang, R. Xie , *et al.*, "Rethinking image aesthetics assessment: Models, datasets and benchmarks." in *IJCAI*, 2022. 55

[476] C. Fu, P. Chen, Y. Shen , *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2024. 57, 58, 68

[477] Y. Liu, H. Duan, Y. Zhang , *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *ECCV*, 2024. 57, 58

[478] P. Xu, W. Shao, K. Zhang , *et al.*, "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models," *IEEE TPAMI*, 2025. 57, 58

[479] Z. Yin, J. Wang, J. Cao , *et al.*, "Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," *arXiv preprint arXiv:2306.06687*, 2023. 58

[480] Q. Ye, H. Xu, G. Xu , *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023. 58

[481] B. Li, R. Wang, G. Wang , *et al.*, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023. 57, 58

[482] B. Li, Y. Ge, Y. Ge , *et al.*, "Seed-bench-2: Benchmarking multimodal large language models," *arXiv preprint arXiv:2311.17092*, 2023. 57, 58, 59, 63, 68

[483] K. Ying, F. Meng, J. Wang , *et al.*, "MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI," in *Proceedings of the 41st ICML*, 2024. 58

[484] X. Fu, Y. Hu, B. Li , *et al.*, "Blink: Multimodal large language models can see but not perceive," *arXiv preprint arXiv:2404.12390*, 2024. 58, 59

[485] L. Chen, J. Li, X. Dong , *et al.*, "Are we on the right way for evaluating large vision-language models?" *arXiv preprint arXiv:2403.20330*, 2024. 58

[486] P. R. Bassi, W. Li, Y. Tang , *et al.*, "Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?" *NeurIPS*, 2024. 58

[487] D. Jang, Y. Cho, S. Lee , *et al.*, "Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation," in *ICLR*, 2025. 58

[488] P. Tong, E. Brown, P. Wu , *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal llms," *NeurIPS*, 2024. 58

[489] K. Cheng, W. Song, J. Fan , *et al.*, "Caparena: Benchmarking and analyzing detailed image captioning in the llm era," *arXiv preprint arXiv:2503.12329*, 2025. 58

[490] K. Narayan, V. VS, and V. M. Patel, "Facexbench: Evaluating multimodal llms on face understanding," *arXiv preprint arXiv:2501.10360*, 2025. 58

[491] F. Wang, X. Fu, J. Y. Huang , *et al.*, "Muirbench: A comprehensive benchmark for robust multi-image understanding," *arXiv preprint arXiv:2406.09411*, 2024. 58

[492] S. Tong, Z. Liu, Y. Zhai , *et al.*, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," *arXiv preprint arXiv:2401.06209*, 2024. 57, 58

[493] M. Cai, H. Liu, D. Park , *et al.*, "Vip-llava: Making large multimodal models understand arbitrary visual prompts," *arXiv preprint arXiv:2312.00784*, 2024. 58

[494] Y.-F. Zhang, H. Zhang, H. Tian , *et al.*, "Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?" *arXiv preprint arXiv:2408.13257*, 2024. 57, 58

[495] Y. Bitton, H. Bansal, J. Hessel , *et al.*, "Visit-bench: A benchmark for vision-language instruction following inspired by real-world use," *arXiv preprint arXiv:2308.06595*, 2023. 57, 58

[496] Y. Shi, Y. Dong, Y. Ding , *et al.*, "Realunify: Do unified models truly benefit from unification? a comprehensive benchmark," *arXiv preprint arXiv:2509.24897*, 2025. 58, 59, 63

[497] W. Yu, Z. Yang, L. Li , *et al.*, "Mm-vet: Evaluating large multimodal models for integrated capabilities," in *ICML*, 2024. 58

[498] X. Yue, Y. Ni, K. Zhang , *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of CVPR*, 2024. 57, 58

[499] Z. He, X. Wu, P. Zhou , *et al.*, "Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning," *arXiv preprint arXiv:2401.14011*, 2024. 57, 58

[500] G. Zhang, X. Du, B. Chen , *et al.*, "Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark," *arXiv preprint arXiv:2401.11944*, 2024. 57, 58

[501] X. Yue, T. Zheng, Y. Ni , *et al.*, "Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark," in *Proceedings of ACL*, 2025. 57, 58

[502] Y. Zhang, S. Lei, R. Qiao , *et al.*, "Multi-dimensional insights: Benchmarking real-world personalization in large multimodal models," *arXiv preprint arXiv:2412.12606*, 2024. 57, 58

[503] Y. Benchekroun, M. Dervishi, M. Ibrahim , *et al.*, "Worldsense: A synthetic benchmark for grounded reasoning in large language models," *arXiv preprint arXiv:2311.15930*, 2023. 58

[504] Z. Cheng, Y. Tu, R. Li , *et al.*, "Embodiedeval: Evaluate multimodal llms as embodied agents," *arXiv preprint arXiv:2501.11858*, 2025. 58

[505] X. Zhao, W. Xu, B. Liu , *et al.*, "Msearth: A benchmark for multimodal scientific comprehension of earth science," *arXiv preprint arXiv:2505.20740*, 2025. 58

[506] X. Han, Q. You, Y. Liu , *et al.*, "Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models," *arXiv preprint arXiv:2311.11567*, 2023. 57, 58

[507] J. Nie, G. Zhang, W. An , *et al.*, "Mmrel: A relation understanding benchmark in the mllm era," *arXiv preprint arXiv:2406.09121*, 2024. 58

[508] H. Shao, S. Qian, H. Xiao , *et al.*, "Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models," *arXiv preprint arXiv:2403.16999*, 2024. 58

[509] Q. Chen, L. Qin, J. Zhang , *et al.*, "M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought," in *Proc. of ACL*, 2024. 57, 58

[510] W. Wang, M. Shi, Q. Li , *et al.*, "The all-seeing project: Towards panoptic visual recognition and understanding of the open world," *arXiv preprint arXiv:2308.01907*, 2023. 58

[511] X. Zhou, J. He, L. Chen , *et al.*, "Miceval: Unveiling multimodal chain of thought's quality via image description and reasoning steps," *arXiv preprint arXiv:2410.14668*, 2024. 58

[512] Y. Wei, S. Fu, W. Jiang , *et al.*, "Gita: Graph to visual and textual integration for vision-language graph reasoning," in *The Thirty-eighth Annual NeurIPS*, 2024. 58

[513] D. Yan, Y. Li, Q.-G. Chen , *et al.*, "Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning," *arXiv preprint arXiv:2503.18533*, 2025. 58

[514] Z. Cheng, Q. Chen, J. Zhang , *et al.*, "Comt: A novel benchmark for chain of multi-modal thought on large vision-language models," *arXiv preprint arXiv:2412.12932*, 2025. 58

[515] A. Singh, V. Natarjan, M. Shah , *et al.*, "Towards vqa models that can read," in *CVPR*, 2019. 58, 59

[516] H. Q. Pham, T. K.-B. Nguyen, Q. Van Nguyen , *et al.*, "Viocrvqa: Novel benchmark dataset and vision reader for visual question answering by understanding vietnamese text in images," *Multimedia Syst.*, 2025. 57, 58

[517] Y. Liu, Z. Li, M. Huang , *et al.*, "Ocrbench: on the hidden mystery of ocr in large multimodal models," *Science China Information Sciences*, 2024. 57, 58, 59

[518] L. Fu, B. Yang, Z. Kuang , *et al.*, "Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning," *arXiv preprint arXiv:2501.00321*, 2024. 57, 58, 59

[519] B. Li, Y. Ge, Y. Chen , *et al.*, "Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension," *arXiv preprint arXiv:2404.16790*, 2024. 58, 59

[520] T. Zhang, S. Wang, L. Li , *et al.*, "Vcr: Visual caption restoration," *arXiv preprint arXiv: 2406.06462*, 2024. 58

[521] M. Mathew, D. Karatzas, and C. V. Jawahar, "Docvqa: A dataset for vqa on document images," *arXiv preprint arXiv:2007.00398*, 2021. 58, 60

[522] M. Mathew, V. Bagal, R. P. Tito , *et al.*, "InfographicVQA," *arXiv preprint arXiv:2104.12756*, 2021. 58, 60

[523] R. Xia, S. Mao, X. Yan , *et al.*, "Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models," *arXiv preprint arXiv:2406.11633*, 2024. 58

[524] Y. Ma, Y. Zang, L. Chen , *et al.*, "Mmlongbench-doc: Benchmarking long-context document understanding with visualizations," *arXiv preprint arXiv:2407.01523*, 2024. 58, 60

[525] Z. Wang, M. Xia, L. He , *et al.*, "Charxiv: Charting gaps in realistic chart understanding in multimodal llms," *arXiv preprint arXiv:2406.18521*, 2024. 58, 60

[526] A. Kembhavi, M. Salvato, E. Kolve , *et al.*, "A diagram is worth a dozen images," in *ECCV*, 2016. 58

[527] R. Tanaka, K. Nishida, and S. Yoshida, "Visualmrc: Machine reading comprehension on document images," in *AAAI*, 2021. 58

[528] R. Chaudhry, S. Shekhar, U. Gupta , *et al.*, "Leaf-qa: Locate, encode & attend for figure question answering," in *WACV*, 2020. 58

[529] S. E. Kahou, V. Michalski, A. Atkinson , *et al.*, "Figureqa: An annotated figure dataset for visual reasoning," *arXiv preprint arXiv:1710.07300*, 2018. 58

[530] J. Roberts, T. Lüddecke, R. Sheikh , *et al.*, "Charting New Territories: Exploring the geographic and geospatial capabilities of multimodal LLMs," *arXiv preprint arXiv:2311.14656*, 2023. 58

[531] F. Liu, X. Wang, W. Yao , *et al.*, "Mmc: Advancing multimodal chart understanding with large-scale instruction tuning," *arXiv preprint arXiv:2311.10774*, 2023. 58, 60

[532] M. Huang, L. Han, X. Zhang , *et al.*, "Evochart: A benchmark and a self-training approach towards real-world chart understanding," *arXiv preprint arXiv:2409.01577*, 2024. 58, 60

[533] X. Wu, J. Yang, L. Chai , *et al.*, "Tablebench: A comprehensive and complex benchmark for table question answering," in *AAAI*, 2025. 58, 60

[534] P. Lu, H. Bansal, T. Xia , *et al.*, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," in *ICLR)*, 2024. 57, 58, 59

[535] K. Wang, J. Pan, W. Shi , *et al.*, "Measuring multimodal mathematical reasoning with math-vision dataset," in *The Thirty-eight NeurIPS Datasets and Benchmarks Track*, 2024. 57, 58, 59

[536] C. He, R. Luo, Y. Bai , *et al.*, "Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems," *arXiv preprint arXiv:2402.14008*, 2024. 58, 59

[537] R. Zhang, D. Jiang, Y. Zhang , *et al.*, "Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?" *arXiv preprint arXiv:2403.14624*, 2024. 57, 58, 59

[538] R. Qiao, Q. Tan, G. Dong , *et al.*, "We-math: Does your large multimodal model achieve human-like mathematical reasoning?" *arXiv preprint arXiv:2407.01284*, 2024. 58

[539] K. Huang, J. Guo, Z. Li , *et al.*, "MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations," *arXiv preprint arXiv:2502.06453*, 2025. 58, 59
[540] Y. Wu, W. Yu, Y. Cheng , *et al.*, "Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models," *arXiv preprint arXiv:2406.09295*, 2024. 58
[541] J. Tang, Q. Liu, Y. Ye , *et al.*, "Mtvqa: Benchmarking multilingual text-centric visual question answering," *arXiv preprint arXiv:2405.11985*, 2024. 58
[542] W. Zhang, M. Aljunied, C. Gao , *et al.*, "M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models," *NeurIPS*, 2023. 58
[543] H. Maryam, L. Fu, J. Song , *et al.*, "Dataset and benchmark for urdu natural scenes text detection, recognition and visual question answering," in *ICDAR*, 2024. 58
[544] F. W. Douamba, J. Song, L. Fu , *et al.*, "The first swahili language scene text detection and recognition dataset," *arXiv preprint arXiv:2405.11437*, 2024. 58
[545] Y. Wang, Y. Liu, F. Yu , *et al.*, "Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation," *arXiv preprint arXiv:2407.01081*, 2024. 58
[546] Y. Li, Y. Du, K. Zhou , *et al.*, "Evaluating object hallucination in large vision-language models," *arXiv preprint arXiv:2305.10355*, 2023. 58, 60
[547] F. Liu, K. Lin, L. Li , *et al.*, "Aligning large multi-modal model with robust instruction tuning," *arXiv preprint arXiv:2306.14565*, 2023. 58
[548] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," in *AAAI Technical Track on NLP*, 2024. 58, 60
[549] J. Wang, Y. Zhou, G. Xu , *et al.*, "Evaluation and analysis of hallucination in large vision-language models," *arXiv preprint arXiv:2308.15126*, 2023. 58, 60
[550] Z. Sun, S. Shen, S. Cao , *et al.*, "Aligning large multimodal models with factually augmented RLHF," in *ACL*, 2024. 58, 60
[551] C. Cui, Y. Zhou, X. Yang , *et al.*, "Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges," *arXiv preprint arXiv:2311.03287*, 2023. 58, 60
[552] S. Yang, R. Sun, and X. Wan, "A new benchmark and reverse validation method for passage-level hallucination detection," in *EMNLP*, 2023. 58
[553] T. Guan, F. Liu, X. Wu , *et al.*, "Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *CVPR*, 2024. 58
[554] J. Wang, Y. Wang, G. Xu , *et al.*, "An llm-free multi-dimensional benchmark for mllms hallucination evaluation," *arXiv preprint arXiv:2311.07397*, 2023. 58
[555] A. Ben-Kish, M. Yanuka, M. Alper , *et al.*, "Mitigating open-vocabulary caption hallucinations," *arXiv preprint arXiv:2312.03631*, 2024. 58, 60
[556] W. Huang, H. Liu, M. Guo , *et al.*, "Visual hallucinations of multimodal large language models," *arXiv preprint arXiv:2402.14683*, 2024. 58
[557] H. Qiu, W. Hu, Z.-Y. Dou , *et al.*, "Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models," *arXiv preprint arXiv:2404.13874*, 2024. 58, 60
[558] B. Yan, J. Zhang, Z. Yuan , *et al.*, "Evaluating the quality of hallucination benchmarks for large vision-language models," *arXiv preprint arXiv:2406.17115*, 2024. 58
[559] M. Wu, J. Ji, O. Huang , *et al.*, "Evaluating and analyzing relationship hallucinations in large vision-language models," in *Proceedings of the 41st ICML*, 2024. 58
[560] K. il Lee, M. Kim, S. Yoon , *et al.*, "Vlind-bench: Measuring language priors in large vision-language models," *arXiv preprint arXiv:2406.08702*, 2025. 58, 60
[561] S. Cha, J. Lee, Y. Lee , *et al.*, "Visually dehallucinative instruction generation," in *ICASSP*, 2024. 58
[562] H. Tu, C. Cui, Z. Wang , *et al.*, "How many unicorns are in this image? a safety evaluation benchmark for vision llms," *arXiv preprint arXiv:2311.16101*, 2023. 57, 58
[563] Y. Zhang, Y. Huang, Y. Sun , *et al.*, "Benchmarking trustworthiness of multimodal large language models: A comprehensive study," *arXiv preprint arXiv:2406.07057*, 2024. 58
[564] W. Wang, X. Liu, K. Gao , *et al.*, "Can't see the forest for the trees: Benchmarking multimodal safety awareness for multimodal llms," *arXiv preprint arXiv:2502.11184*, 2025. 58
[565] X. Li, H. Zhou, R. Wang , *et al.*, "Mossbench: Is your multimodal language model oversensitive to safe queries?" *arXiv preprint arXiv:2406.17806*, 2024. 58, 60
[566] S. Wang, X. Cao, J. Zhang , *et al.*, "Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model," *arXiv preprint arXiv:2406.14194*, 2024. 58, 60
[567] C. Fu, Y. Dai, Y. Luo , *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv preprint arXiv:2405.21075*, 2024. 57, 58, 60
[568] K. Li, Y. Wang, Y. He , *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024. 58
[569] J. Zhou, Y. Shu, B. Zhao , *et al.*, "Mlvu: Benchmarking multi-task long video understanding," in *CVPR*, 2025. 58
[570] X. Fang, K. Mao, H. Duan , *et al.*, "Mmbench-video: A long-form multi-shot benchmark for holistic video understanding," *arXiv preprint arXiv:2406.14515*, 2024. 58, 60
[571] Z. Zhao, H. Lu, Y. Huo , *et al.*, "Needle in a video haystack: A scalable synthetic framework for benchmarking video mllms," *arXiv preprint arXiv:2406.09367*, 2024. 58, 60
[572] Y. Liu, S. Li, Y. Liu , *et al.*, "Tempcompass: Do video llms really understand videos?" *arXiv preprint arXiv: 2403.00476*, 2024. 58, 60
[573] B. Li, Y. Wu, Y. Lu , *et al.*, "Veu-bench: Towards comprehensive understanding of video editing," *arXiv preprint arXiv:2504.17828*, 2025. 58, 60
[574] Z. Yu, D. Xu, J. Yu , *et al.*, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *AAAI*, 2019. 58, 60
[575] Y. Zhao, Y. Zeng, Y. Qi , *et al.*, "V2p-bench: Evaluating video-language understanding with visual prompts for better human-model interaction," *arXiv preprint arXiv:2503.17736*, 2025. 58, 60
[576] W. Hong*, Y. Cheng*, Z. Yang* , *et al.*, "Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models," *arXiv preprint arXiv:2501.02955*, 2024. 58, 60
[577] Q. Wu, Q. Zheng, Y. Zhang , *et al.*, "H2vu-benchmark: A comprehensive benchmark for hierarchical holistic video understanding," *arXiv preprint arXiv:2503.24008*, 2025. 58, 60
[578] C. Tu, L. Zhang, P. Chen , *et al.*, "Favor-bench: A comprehensive benchmark for fine-grained video motion understanding," *arXiv preprint arXiv:2503.14935*, 2025. 58, 60
[579] D. Saravanan, V. Gupta, D. Singh , *et al.*, "VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment," in *CVPR*, 2025. 58
[580] Y. Jang, Y. Song, C. D. Kim , *et al.*, "Video Question Answering with Spatio-Temporal Reasoning," *IJCV*, 2019. 58, 60
[581] L. Jang, Y. Li, D. Zhao , *et al.*, "Videowebarena: Evaluating long context multimodal agents with video understanding web tasks," *arXiv preprint arXiv:2410.19100*, 2024. 58
[582] Z. Yang, Y. Hu, Z. Du , *et al.*, "Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding," *arXiv preprint arXiv:2502.10810*, 2025. 58, 61
[583] G. Chen, Y. Liu, Y. Huang , *et al.*, "Cg-bench: Clue-grounded question answering benchmark for long video understanding," *arXiv preprint arXiv:2412.12075*, 2024. 58, 61
[584] A. Nagrani, S. Menon, A. Iscen , *et al.*, "Minerva: Evaluating complex video reasoning," *arXiv preprint arXiv:2505.00681*, 2025. 58, 61
[585] R. Wu, X. Ma, H. Ci , *et al.*, "Longvitu: Instruction tuning for long-form video understanding," *arXiv preprint arXiv:2501.05037*, 2025. 58, 61
[586] W. Zhou, K. Cao, H. Zheng , *et al.*, "X-lebench: A benchmark for extremely long egocentric video understanding," *arXiv preprint arXiv:2501.06835*, 2025. 58, 61
[587] H. Zou, T. Luo, G. Xie , *et al.*, "Hlv-1k: A large-scale hour-long video benchmark for time-specific long video understanding," *arXiv preprint arXiv:2501.01645*, 2025. 58, 60
[588] W. Wang, Z. He, W. Hong , *et al.*, "Lvbench: An extreme long video understanding benchmark," *arXiv preprint arXiv:2406.08035*, 2024. 58, 60
[589] Y. Du, K. Zhou, Y. Huo , *et al.*, "Towards event-oriented long video understanding," *arXiv preprint arXiv:2406.14129*, 2024. 58, 61
[590] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," *arXiv preprint arXiv:2308.09126*, 2023. 58, 61

[591] E. Song, W. Chai, W. Xu , *et al.*, "Video-mmlu: A massive multi-discipline lecture understanding benchmark," *arXiv preprint arXiv:2504.14693*, 2025. 58, 61

[592] K. Hu, P. Wu, F. Pu , *et al.*, "Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos," *arXiv preprint arXiv:2501.13826*, 2025. 58

[593] Y. Li, Y. Zhang, T. Lin , *et al.*, "Sti-bench: Are mllms ready for precise spatial-temporal world understanding?" *arXiv preprint arXiv:2503.23765*, 2025. 58, 61

[594] B. Zhao, J. Fang, Z. Dai , *et al.*, "Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces," *arXiv preprint arXiv:2503.06157*, 2025. 58, 61

[595] M. Cao, P. Hu, Y. Wang , *et al.*, "Video simpleqa: Towards factuality evaluation in large video language models," *arXiv preprint arXiv:2503.18923*, 2025. 58, 61

[596] Z. Cheng, J. Hu, Z. Liu , *et al.*, "V-star: Benchmarking video-llms on video spatio-temporal reasoning," *arXiv preprint arXiv:2503.11495*, 2025. 58

[597] H. Wei, Y. Yuan, X. Lan , *et al.*, "Instructionbench: An instructional video understanding benchmark," *arXiv preprint arXiv:2504.05040*, 2025. 58, 61

[598] Y. Li, J. Niu, Z. Miao , *et al.*, "Ovo-bench: How far is your video-llms from real-world online video understanding?" *arXiv preprint arXiv:2501.05510*, 2025. 58, 61

[599] C. Li, Q. Chen, Z. Li , *et al.*, "Vcbench: A controllable benchmark for symbolic and abstract challenges in video cognition," *arXiv preprint arXiv:2411.09105*, 2024. 58, 61

[600] W. Xu, J. Wang, W. Wang , *et al.*, "Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models," *arXiv preprint arXiv:2504.15279*, 2025. 58

[601] A. Agarwal, S. Panda, A. Charles , *et al.*, "Mvtamperbench: Evaluating robustness of vision-language models," *arXiv preprint arXiv:2412.19794*, 2024. 58, 61

[602] Y. Yang, Y. Guo, H. Lu , *et al.*, "Vidlbeval: Benchmarking and mitigating language bias in video-involved lvlms," *arXiv preprint arXiv:2502.16602*, 2025. 58, 61

[603] M. Kong, X. Zeng, L. Chen , *et al.*, "Mhbench: Demystifying motion hallucination in videollms," *AAAI*, 2025. 58

[604] W. Y. Choong, Y. Guo, and M. Kankanhalli, "Vidhal: Benchmarking temporal hallucinations in vision llms," *arXiv preprint arXiv:2411.16771*, 2024. 58, 61

[605] Y. Wang, Y. Wang, D. Zhao , *et al.*, "Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models," *arXiv preprint arXiv:2406.16338*, 2024. 58, 61

[606] S. Zhou, J. Xiao, Q. Li , *et al.*, "Egotextvqa: Towards egocentric scene-text aware video question answering," *arXiv preprint arXiv:2502.07411*, 2025. 58

[607] Y. Shi, H. Wang, W. Xie , *et al.*, "Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios," *arXiv preprint arXiv:2505.21333*, 2025. 58, 61

[608] Y. Fei, Y. Gao, X. Xian , *et al.*, "Do current video llms have strong ocr abilities? a preliminary study," *arXiv preprint arXiv:2412.20613*, 2024. 58, 61

[609] G. Yin, H. Bai, S. Ma , *et al.*, "MMAU: A holistic benchmark of agent capabilities across diverse domains," in *NAACL*, 2025. 57, 58, 61

[610] G. Maimon, A. Roth, and Y. Adi, "Salmon: A suite for acoustic language model evaluation," in *ICASSP*, 2025. 58, 61

[611] Q. Yang, J. Xu, W. Liu , *et al.*, "AIR-bench: Benchmarking large audio-language models via generative comprehension," in *ACL (Volume 1: Long Papers)*, 2024. 57, 58, 61

[612] B. Wang, X. Zou, G. Lin , *et al.*, "AudioBench: A universal benchmark for audio large language models," in *NAACL on Human Language Technologies (Volume 1: Long Papers)*, 2025. 57, 58, 61

[613] B. Weck, I. Manco, E. Benetos , *et al.*, "Muchomusic: Evaluating music understanding in multimodal audio-language models," in *ISMIR*, 2024. 57, 58, 61

[614] Y. Chen, X. Yue, C. Zhang , *et al.*, "Voicebench: Benchmarking llm-based voice assistants," *arXiv preprint arXiv:2410.17196*, 2024. 58, 61

[615] S. Deshmukh, S. Han, H. Bukhari , *et al.*, "Audio entailment: Assessing deductive reasoning for audio understanding," *arXiv preprint arXiv:2407.18062*, 2024. 58, 61

[616] W. Cui, X. Jiao, Z. Meng , *et al.*, "Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models," *arXiv preprint arXiv:2501.04962*, 2025. 58

[617] A. S. Penamakuri, K. Chhatre, and A. Jain, "Audiopedia: Audio qa with knowledge," *arXiv preprint arXiv:2412.20619*, 2024. 58, 61

[618] Y. Yang, J. Zhuang, G. Sun , *et al.*, "Acvubench: Audio-centric video understanding benchmark," *arXiv preprint arXiv:2503.19951*, 2025. 58, 61

[619] L. Xie, G. Z. Wei, A. Kuthiala , *et al.*, "Maverix: Multimodal audio-visual evaluation reasoning index," *arXiv preprint arXiv:2503.21699*, 2025. 58, 62

[620] W. Xie, Y.-F. Zhang, C. Fu , *et al.*, "Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models," *arXiv preprint arXiv:2504.03641*, 2025. 58, 59, 63, 68

[621] J. Wang, H. Duan, Y. Zhao , *et al.*, "Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms," *arXiv preprint arXiv:2504.08358*, 2025. 58, 59

[622] X. Hu, R. Wang, Y. Fang , *et al.*, "Ella: Equip diffusion models with llm for enhanced semantic alignment," *arXiv preprint arXiv:2403.05135*, 2024. 58, 59

[623] X. Wu, D. Yu, Y. Huang , *et al.*, "Conceptmix: A compositional image generation benchmark with controllable difficulty," *arXiv preprint arXiv:2408.14339*, 2024. 58, 59, 62

[624] N. Ruiz, Y. Li, V. Jampani , *et al.*, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *arXiv preprint arXiv:2208.12242*, 2022. 58, 59, 62

[625] Y. Kirstain, A. Polyak, U. Singer , *et al.*, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *arXiv preprint arXiv:2305.01569*, 2023. 58, 59, 62

[626] Y. Tian, Y. Li, B. Chen , *et al.*, "Ai-generated image quality assessment in visual communication," *arXiv preprint arXiv:2412.15677*, 2024. 58

[627] K. Huang, K. Sun, E. Xie , *et al.*, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *NeurIPS*, 2023. 58, 62

[628] Y. Hu, B. Liu, J. Kasai , *et al.*, "Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering," *arXiv preprint arXiv:2303.11897*, 2023. 58, 59, 62

[629] J. Cho, Y. Hu, J. Baldridge , *et al.*, "Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation," in *ICLR*, 2024. 58, 59

[630] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *arXiv preprint arXiv:2310.11513*, 2023. 58, 59, 68

[631] J. Yu, Y. Xu, J. Y. Koh , *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, 2022. 58, 59

[632] K. Huang, C. Duan, K. Sun , *et al.*, " T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation ," *TPAMI*, 5555. 58, 62

[633] E. M. Bakr, P. Sun, X. Shen , *et al.*, "Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models," in *ICCV*, 2023. 58

[634] Y. Lim, H. Choi, and H. Shim, "Evaluating image hallucination in text-to-image generation with question-answering," in *AAAI*, 2025. 58, 62

[635] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers," *arXiv preprint arXiv:2202.04053*, 2022. 58, 62

[636] Y. Huang, C. Gao, S. Wu , *et al.*, "On the trustworthiness of generative foundation models: Guideline, assessment, and perspective," *arXiv preprint arXiv:2502.14296*, 2025. 58, 62

[637] H. Luo, H. Huang, Z. Deng , *et al.*, "Bigbench: A unified benchmark for evaluating multi-dimensional social biases in text-to-image models," *arXiv preprint arXiv:2407.15240*, 2025. 58

[638] H. Luo, Z. Deng, R. Chen , *et al.*, "Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models," *arXiv preprint arXiv:2405.17814*, 2025. 58, 62

[639] L. Li, Z. Shi, X. Hu , *et al.*, "T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation," *arXiv preprint arXiv:2501.12612*, 2025. 58, 62

[640] T. Lee, M. Yasunaga, C. Meng , *et al.*, "Holistic evaluation of text-to-image models," *arXiv preprint arXiv:2311.04287*, 2023. 58, 62

[641] Y. Niu, M. Ning, M. Zheng , *et al.*, "Wise: A world knowledge-informed semantic evaluation for text-to-image generation," *arXiv preprint arXiv:2503.07265*, 2025. 58, 59, 62, 68

[642] X. Fu, M. He, Y. Lu , *et al.*, "Commonsense-t2i challenge: Can text-to-image generation models understand commonsense?" *arXiv preprint arXiv:2406.07546*, 2024. 58, 59, 62

[643] F. Meng, W. Shao, L. Luo , *et al.*, "Phybench: A physical commonsense benchmark for evaluating text-to-image models," *arXiv preprint arXiv:2406.11802*, 2024. 58, 59

[644] S. Sheynin, A. Polyak, U. Singer , *et al.*, "Emu edit: Precise image editing via recognition and generation tasks," in *CVPR*, 2024. 58, 59, 62

[645] T. Li, M. Ku, C. Wei , *et al.*, "Dreamedit: Subject-driven image editing," *arXiv preprint arXiv:2306.12624*, 2023. 58, 59

[646] X. Ju, A. Zeng, Y. Bian , *et al.*, "Pnp inversion: Boosting diffusion-based editing with 3 lines of code," *ICLR*, 2024. 58, 59, 62

[647] S. Liu, Y. Han, P. Xing , *et al.*, "Step1x-edit: A practical framework for general image editing," *arXiv preprint arXiv:2504.17761*, 2025. 58, 59

[648] Y. Wu, Z. Li, X. Hu , *et al.*, "Kris-bench: Benchmarking next-level intelligent image editing models," *arXiv preprint arXiv:2505.16707*, 2025. 58

[649] D. Chang, M. Cao, Y. Shi , *et al.*, "Bytemorph: Benchmarking instruction-guided image editing with non-rigid motions," *arXiv preprint arXiv:2506.03107*, 2025. 58

[650] Q. Li, Z. Xing, R. Wang , *et al.*, "Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance," *arXiv preprint arXiv:2503.16421*, 2025. 58, 59

[651] X. Zhao, P. Zhang, K. Tang , *et al.*, "Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing," *arXiv preprint arXiv:2504.02826*, 2025. 58, 59, 62

[652] B. Krojer, D. Vattikonda, L. Lara , *et al.*, "Learning action and reasoning-centric image editing from videos and simulations," *arXiv preprint arXiv:2407.03471*, 2024. 58, 62

[653] T. Xiong, Y. Wu, E. Xie , *et al.*, "Editing massive concepts in text-to-image diffusion models," *arXiv preprint arXiv:2403.13807*, 2024. 58, 59, 62

[654] K. Guan, Z. Lai, Y. Sun , *et al.*, "Etva: Evaluation of text-to-video alignment via fine-grained question generation and answering," *arXiv preprint arXiv:2503.16867*, 2025. 58, 59

[655] H. Han, S. Li, J. Chen , *et al.*, "Video-bench: Human-aligned video generation benchmark," *arXiv preprint arXiv:2504.04907*, 2025. 58, 59

[656] Y. Liu, X. Cun, X. Liu , *et al.*, "Evalcrafter: Benchmarking and evaluating large video generation models," in *CVPR*, 2024. 58, 59, 62

[657] X. Ju, Y. Gao, Z. Zhang , *et al.*, "Miradata: A large-scale video dataset with long durations and structured captions," *arXiv preprint arXiv:2407.06358*, 2024. 58, 62

[658] T. Kou, X. Liu, Z. Zhang , *et al.*, "Subjective-aligned dateset and metric for text-to-video quality assessment," *arXiv preprint arXiv:2403.11956*, 2024. 58

[659] Z. Chen, W. Sun, Y. Tian , *et al.*, "Gaia: Rethinking action quality assessment for ai-generated videos," *NeurIPS*, 2024. 58

[660] K. Sun, K. Huang, X. Liu , *et al.*, "T2v-compbench: A comprehensive benchmark for compositional text-to-video generation," *arXiv preprint arXiv:2407.14505*, 2024. 58

[661] P. Ji, C. Xiao, H. Tai , *et al.*, "T2vbench: Benchmarking temporal dynamics for text-to-video generation," in *CVPR Workshops*, 2024. 58

[662] R. Szeto and J. J. Corso, "The devil is in the details: A diagnostic evaluation benchmark for video inpainting," in *CVPR*, 2022. 58

[663] Z. Zhang, W. Sun, X. Li , *et al.*, "Human-activity agv quality assessment: A benchmark dataset and an objective evaluation metric," *arXiv preprint arXiv:2411.16619*, 2025. 58

[664] H. Wu, E. Zhang, L. Liao , *et al.*, "Towards explainable video quality assessment: A database and a language-prompted approach," in *ACM MM*, 2023. 58, 59

[665] X. He, D. Jiang, G. Zhang , *et al.*, "Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation," *arXiv preprint arXiv:2406.15252*, 2024. 58, 59, 62

[666] Z. Zhang, W. Sun, X. Li , *et al.*, "Benchmarking multi-dimensional aigc video quality assessment: A dataset and unified model," *arXiv preprint arXiv:2407.21408*, 2024. 58, 59

[667] Z. Huang, Y. He, J. Yu , *et al.*, "VBench: Comprehensive benchmark suite for video generative models," in *CVPR*, 2024. 58, 59

[668] Z. Huang, F. Zhang, X. Xu , *et al.*, "Vbench++: Comprehensive and versatile benchmark suite for video generative models," *arXiv preprint arXiv:2411.13503*, 2024. 58

[669] X. Guo, J. Huo, Z. Shi , *et al.*, "T2vtextbench: A human evaluation benchmark for textual control in video generation models," *arXiv preprint arXiv:2505.04946*, 2025. 58

[670] S. Yuan, J. Huang, Y. Xu , *et al.*, "Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation," *NeurIPS*, 2024. 58, 59, 62

[671] Y. Qin, Z. Shi, J. Yu , *et al.*, "Worldsimbench: Towards video generation models as world simulators," *arXiv preprint arXiv:2410.18072*, 2024. 58

[672] F. Meng, J. Liao, X. Tan , *et al.*, "Towards world simulator: Crafting physical commonsense-based benchmark for video generation," *arXiv preprint arXiv:2410.05363*, 2024. 58, 59, 62

[673] H. Duan, H.-X. Yu, S. Chen , *et al.*, "Worldscore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025. 58

[674] C. Zhang, D. Cherniavskii, A. Zadaianchuk , *et al.*, "Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments," *arXiv preprint arXiv:2504.02918*, 2025. 58, 62

[675] Y. Wang, X. He, K. Wang , *et al.*, "Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation," *arXiv preprint arXiv:2412.16211*, 2024. 58, 62

[676] X. Guo, J. Huo, Z. Shi , *et al.*, "T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation," *arXiv preprint arXiv:2505.00337*, 2025. 58

[677] J. Gu, X. Liu, Y. Zeng , *et al.*, ""phyworldbench": A comprehensive evaluation of physical realism in text-to-video models," *arXiv preprint arXiv:2507.13428*, 2025. 58

[678] Y. Miao, Y. Zhu, Y. Dong , *et al.*, "T2vsafetybench: Evaluating the safety of text-to-video generative models," *arXiv preprint arXiv:2407.05965*, 2024. 58, 62

[679] S. Sun, X. Liang, S. Fan , *et al.*, "Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment," *arXiv preprint arXiv:2408.11481*, 2024. 58, 59, 62

[680] Y. Chen, P. Chen, X. Zhang , *et al.*, "Editboard: Towards a comprehensive evaluation benchmark for text-based video editing models," *arXiv preprint arXiv:2409.09668*, 2025. 58, 62

[681] J. Z. Wu, X. Li, D. Gao , *et al.*, "Cvpr 2023 text guided video editing competition," *arXiv preprint arXiv:2310.16003*, 2023. 58, 63

[682] F. Fan, C. Luo, W. Gao , *et al.*, "Aigcbench: Comprehensive evaluation of image-to-video content generated by ai," *TBench*, 2024. 58, 59, 63

[683] W. Wang and Y. Yang, "Tip-i2v: A million-scale real text and image prompt dataset for image-to-video generation," *arXiv preprint arXiv:2411.04709*, 2024. 58, 59, 63

[684] W. Ren, H. Yang, G. Zhang , *et al.*, "Consisti2v: Enhancing visual consistency for image-to-video generation," *arXiv preprint arXiv:2402.04324*, 2024. 58, 63

[685] F. Jiang, Z. Lin, F. Bu , *et al.*, "S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information," *arXiv preprint arXiv:2503.05085*, 2025. 58, 63

[686] M. de Seyssel, A. D'Avirro, A. Williams , *et al.*, "Emphassess : a prosodic benchmark on assessing emphasis transfer in speech-to-speech models," *arXiv preprint arXiv:2312.14069*, 2023. 58, 63

[687] C. Minixhofer, O. Klejch, and P. Bell, "Ttsds-text-to-speech distribution score," in *SLT*, 2024. 58, 63

[688] Y. Pan, X. He, C. Mao , *et al.*, "Ice-bench: A unified and comprehensive benchmark for image creating and editing," *arXiv preprint arXiv:2503.14482*, 2025. 58, 63

[689] M. Ku, T. Li, K. Zhang , *et al.*, "Imagenhub: Standardizing the evaluation of conditional image generation models," in *ICLR*, 2024. 58, 59, 63

[690] W. Feng, J. Li, M. Saxon , *et al.*, "Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation," *arXiv preprint arXiv:2406.08656*, 2024. 58, 63

[691] Y. Yang, K. Fan, S. Sun , *et al.*, "Videogen-eval: Agent-based system for video generation evaluation," *arXiv preprint arXiv:2503.23452*, 2025. 58, 59

[692] D. Jiang, M. Ku, T. Li , *et al.*, "Genai arena: An open evaluation platform for generative models," *arXiv preprint arXiv:2406.04485*, 2024. 58, 59, 63

[693] M. Liu, Z. Xu, Z. Lin , *et al.*, "Holistic evaluation for interleaved text-and-image generation," *arXiv preprint arXiv:2406.14643*, 2024. 58, 59, 63, 68

[694] P. Xia, S. Han, S. Qiu , *et al.*, "Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models," *arXiv preprint arXiv:2410.10139*, 2025. 58, 59, 63, 68

[695] Y. Li, H. Wang, Q. Zhang , *et al.*, "Unieval: Unified holistic evaluation for unified multimodal understanding and generation," *arXiv preprint arXiv:2505.10483*, 2025. 58, 59, 63, 68

[696] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," *arXiv preprint arXiv:2202.04053*, 2023. 59

[697] W. Wang and Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models," *arXiv preprint arXiv:2403.06098*, 2024. 59

[698] J. Pont-Tuset, F. Perazzi, S. Caelles , *et al.*, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2018. 59

[699] L. Mehl, J. Schmalfuss, A. Jahedi , *et al.*, "Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo," *arXiv preprint arXiv:2303.01943*, 2023. 59

[700] R. Feng, W. Weng, Y. Wang , *et al.*, "Ccedit: Creative and controllable video editing via diffusion models," *arXiv preprint arXiv:2309.16496*, 2024. 59

[701] R. Krishna, K. Hata, F. Ren , *et al.*, "Dense-captioning events in videos," *arXiv preprint arXiv:1705.00754*, 2017. 59

[702] N. Kumari, B. Zhang, R. Zhang , *et al.*, "Multi-concept customization of text-to-image diffusion," *arXiv preprint arXiv:2212.04488*, 2023. 59

[703] D. Chen, R. Chen, S. Pu , *et al.*, "Interleaved scene graph for interleaved text-and-image generation assessment," *arXiv preprint arXiv:2411.17188*, 2024. 59, 63

[704] M. Kazemi, N. Dikkala, A. Anand , *et al.*, "Remi: A dataset for reasoning with multiple images," *arXiv preprint arXiv:2406.09175*, 2024. 59

[705] C. Liu, H. Wu, Y. Zhong , *et al.*, "Intelligent grimm - open-ended visual storytelling via latent diffusion models," in *CVPR*, 2024. 59

[706] Corran, "Pexel videos," 2022. [Online]. Available: https://huggingface.co/datasets/Corran/pexelvideos 59

[707] W. Wang, L. Ding, M. Zeng , *et al.*, "Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models," in *AAAI*, 2025. 59

[708] W. Ye, G. Zheng, Y. Ma , *et al.*, "Mm-spubench: Towards better understanding of spurious biases in multimodal llms," *arXiv preprint arXiv:2406.17126*, 2024. 60

[709] A. Chinchure, P. Shukla, G. Bhatt , *et al.*, "Tibet: Identifying and evaluating biases in text-to-image generative models," *arXiv preprint arXiv:2312.01261*, 2023. 62

[710] A. Brohan, N. Brown, J. Carbajal , *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *arXiv preprint arXiv:2212.06817*, 2022. 64

[711] C. Chi, Z. Xu, S. Feng , *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2023. 64

[712] K. Black, N. Brown, D. Driess , *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024. 64

[713] Y. Shentu, P. Wu, A. Rajeswaran , *et al.*, "From llms to actions: Latent codes as bridges in hierarchical robot control," *arXiv preprint arXiv:2405.04798*, 2025. 64

[714] J. Wen, Y. Zhu, Z. Tang , *et al.*, "Dexvla: Vision-language model with plug-in diffusion expert for visuomotor policy learning," *arXiv preprint arXiv:2502.05855*, 2025. 64

[715] J. Li, Y. Zhu, Z. Tang , *et al.*, "Coa-vla: Improving vision-language-action models via visual-textual chain-of-affordance," *arXiv preprint arXiv:2412.20451*, 2024. 64

[716] Y. Tian, S. Yang, J. Zeng , *et al.*, "Predictive inverse dynamics models are scalable learners for robotic manipulation," *arXiv preprint arXiv:2412.15109*, 2024. 64

[717] Q. Zhao, Y. Lu, M. J. Kim , *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," *arXiv preprint arXiv:2503.22020*, 2025. 64

[718] W. Zhang, H. Liu, Z. Qi , *et al.*, "Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge," *arXiv preprint arXiv:2507.04447*, 2025. 64

[719] C.-L. Cheang, G. Chen, Y. Jing , *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024. 64

[720] S. Li, Y. Gao, D. Sadigh , *et al.*, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025. 64

[721] C. Zhu, R. Yu, S. Feng , *et al.*, "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," *arXiv preprint arXiv:2504.02792*, 2025. 64

[722] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020. 64

[723] L. Chen, P. Wu, K. Chitta , *et al.*, "End-to-end autonomous driving: Challenges and frontiers," *IEEE TPAMI*, 2024. 64

[724] Y. Hu, J. Yang, L. Chen , *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023. 64

[725] J.-J. Hwang, R. Xu, H. Lin , *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024. 64

[726] Y. Chen, Y. Wang, and Z. Zhang, "Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers," *arXiv preprint arXiv:2412.18607*, 2024. 64

[727] K. Zhang, Z. Tang, X. Hu , *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," *arXiv preprint arXiv:2506.24113*, 2025. 64

[728] Y. Wu, H. Zhang, T. Lin , *et al.*, "Generating multimodal driving scenes via next-scene prediction," *arXiv preprint arXiv:2503.14945*, 2025. 64

[729] F. Jia, W. Mao, Y. Liu , *et al.*, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023. 64

[730] S. Zeng, X. Chang, M. Xie , *et al.*, "Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving," *arXiv preprint arXiv:2505.17685*, 2025. 64

[731] X. Zhou, D. Liang, S. Tu , *et al.*, "Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation," *arXiv preprint arXiv:2501.14729*, 2025. 64

[732] J. Wei, S. Yuan, P. Li , *et al.*, "Occllama: An occupancy-language-action generative world model for autonomous driving," *arXiv preprint arXiv:2409.03272*, 2024. 64

[733] T. Xu, H. Lu, X. Yan , *et al.*, "Occ-llm: Enhancing autonomous driving with occupancy-based large language models," *arXiv preprint arXiv:2502.06419*, 2025. 64

[734] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, 2022. 64

[735] N. Agarwal, A. Ali, M. Bala , *et al.*, "Cosmos world foundation model platform for physical ai," *arXiv preprint arXiv:2501.03575*, 2025. 65, 68

[736] A. Team, H. Zhu, Y. Wang , *et al.*, "Aether: Geometric-aware unified world modeling," *arXiv preprint arXiv:2503.18945*, 2025. 65

[737] J. Chen, H. Zhu, X. He , *et al.*, "Deepverse: 4d autoregressive video generation as a world model," *arXiv preprint arXiv:2506.01103*, 2025. 65

[738] H. Zhen, Q. Sun, H. Zhang , *et al.*, "Tesseract: learning 4d embodied world models," *arXiv preprint arXiv:2504.20995*, 2025. 65

[739] M. Hassan, S. Stapf, A. Rahimi , *et al.*, "Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control," in *CVPR*, 2025. 65

[740] J. Guo, Y. Ding, X. Chen , *et al.*, "Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation," *arXiv preprint arXiv:2503.15208*, 2025. 65

[741] J. Ye and H. Tang, "Multimodal large language models for medicine: A comprehensive survey," *arXiv preprint arXiv:2504.21051*, 2025. 65

[742] J. Wu, M. Zhong, S. Xing , *et al.*, "Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks," *NeurIPS*, 2024. 65

[743] A. Kirillov, E. Mintun, N. Ravi , *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 65

[744] Y. Li, R. Hou, H. Chang , *et al.*, "Unipose: A unified multimodal framework for human pose comprehension, generation and editing," *arXiv preprint arXiv:2411.16781*, 2024. 65

[745] X. Zou, J. Yang, H. Zhang , *et al.*, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023. 65

[746] Y. Xu, Z. He, M. Kan , *et al.*, "Jodi: Unification of visual generation and understanding via joint modeling," *arXiv preprint arXiv:2505.19084*, 2025. 65

[747] J. Ye, Z. Wang, R. Zhao , *et al.*, "Shapellm-omni: A native multimodal llm for 3d generation and understanding," *arXiv preprint arXiv:2506.01853*, 2025. 65
[748] J. Wang, M. Chen, N. Karaev , *et al.*, "Vggt: Visual geometry grounded transformer," in *CVPR*, 2025. 65
[749] N. Team, C. Han, G. Li , *et al.*, "Nextstep-1: Toward autoregressive image generation with continuous tokens at scale," *arXiv preprint arXiv:2508.10711*, 2025. 66
[750] M. Huh, B. Cheung, T. Wang , *et al.*, "The platonic representation hypothesis," in *ICML*, 2024. 67
[751] J. Wang, Y. Jiang, Z. Yuan , *et al.*, "Omnitokenizer: A joint image-video tokenizer for visual generation," *NeurIPS*, 2024. 68
[752] Z. Yang, J. Teng, W. Zheng , *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024. 68
[753] Z. Yang, Y. Zhang, F. Meng , *et al.*, "Teal: Tokenize and embed all for multi-modal large language models," *arXiv preprint arXiv:2311.04589*, 2024. 68
[754] Y. Zhang, K. Gong, K. Zhang , *et al.*, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023. 68