Next Token Prediction Towards Multimodal Intelligence: A Comprehensive Survey

LIANG CHEN[1†] ZEKUN WANG[2*] SHUHUAI REN[1*] LEI LI[3*] HAOZHE ZHAO[1*] YUNSHUI LI[4*] ZEFAN CAI[1] HONGCHENG GUO[2] LEI ZHANG[4] YIZHE XIONG[5] YICHI ZHANG[1] RUOYU WU[1] QINGXIU DONG[1] GE ZHANG[6] JIAN YANG[8] LINGWEI MENG[7] SHUJIE HU[7] YULONG CHEN[9] JUNYANG LIN[8] SHUAI BAI[8] ANDREAS VLACHOS[9] XU TAN[10] MINJIA ZHANG[11] WEN XIAO[10] AARON YEE[12,13] TIANYU LIU[8] BAOBAO CHANG[1]

Abstract: Building on the foundations of language modeling in natural language processing, Next Token Prediction (NTP) has evolved into a versatile training objective for machine learning tasks across various modalities, achieving considerable success. As Large Language Models (LLMs) have advanced to unify understanding and generation tasks within the textual modality, recent research has shown that tasks from different modalities can also be effectively encapsulated within the NTP framework, transforming the multimodal information into tokens and predict the next one given the context. This survey introduces a comprehensive taxonomy that unifies both understanding and generation within multimodal learning through the lens of NTP. The proposed taxonomy covers five key aspects: Multimodal tokenization, MMNTP model architectures, unified task representation, datasets & evaluation, and open challenges. This new taxonomy aims to aid researchers in their exploration of multimodal intelligence. An associated GitHub repository collecting the latest papers and repos is available at https://github.com/LMM101/Awesome-Multimodal-Next-Token-Prediction.

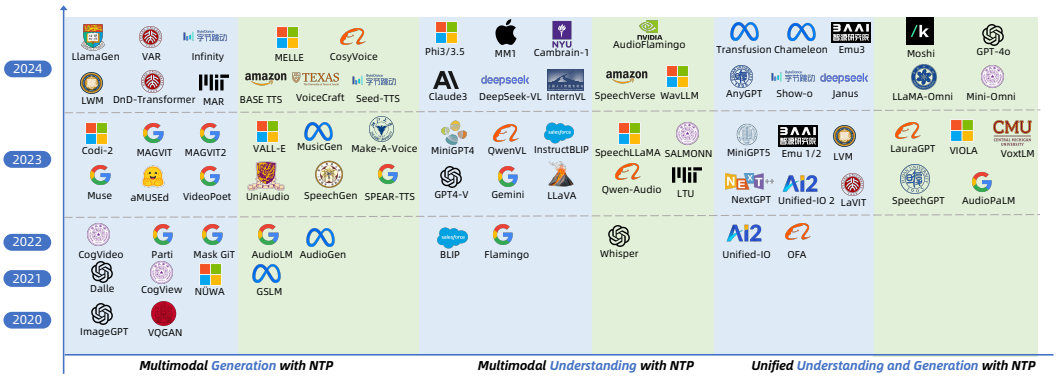

Fig. 1. Historical development of LMMs utilizing Next-Token Prediction. Models featuring vision and more modalities are set in blue background while models that support audio modality are set in green backgrounds.

1 INTRODUCTION

Humans' engagement with the universe is a tapestry, interwoven with the threads of various modalities. Humans can see and sketch paintings, read and write epics, listen and compose music, touch and sculpture heroes, ponder and make movements. These modalities – specific information types such as vision, sound, and language – are the channels through which humans interpret and

Authors' affiliations: [1]Peking University [2]Beihang Univerisy [3]University of Hongkong [4]Shenzhen Institute of Advanced Technology, China Academy of Sciences [5]Tsinghua University [6]M-A-P [7]The Chinese University of Hong Kong [8]Alibaba Group [9]University of Cambridge [10]Microsoft Research [11]UIUC [12]Humanify Inc. [13]Zhejiang University
Authors' contributions: [†] project lead. [*] core contributors. Full contributions are in Section 7.
Corresponding to: Liang Chen <leo.liang.chen@outlook.com>, Baobao Chang <chbb@pku.edu.cn>

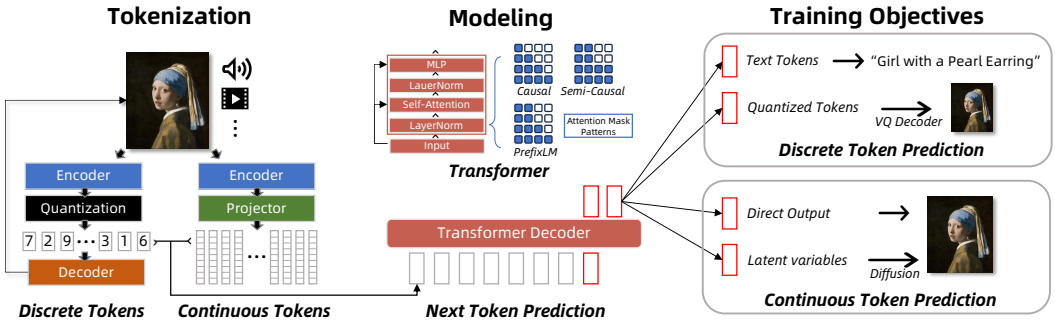

Fig. 2. General pipeline of Multimodal Learning with Next Token Prediction (MMNTP).

respond to the world. This multifaceted interaction highlights the intertwined nature of perception and response in human experience. As a specialized area within Artificial Intelligence (AI) research, multimodal Learning focuses on creating systems capable of understanding and generating various multimodal information [16].

A paradigm shift has emerged in the field of AI across multiple modalities, transitioning from specialized unimodal models trained for a single task to versatile multimodal ones dealing with a diverse array of tasks [150]. This shift is largely attributed to the advancement of Large Language Models (LLMs) in the Natural Language Processing (NLP) field such as GPT-3 [34], ChatGPT [300] and LLaMA [378], which unify multiple natural language understanding and generation tasks with a single Next Token Prediction (NTP) objective. The original task of NTP is to predict the next token (which can be a word, subword, or character) in a given sequence of text based on the context provided by preceding tokens. The NTP paradigm has been proven to be scalable given abundant data and computational resources in the lens of scaling law research [192, 472].

Simultaneously, researchers have explored the incorporation of non-textual input and output modalities into large language models, sparking interest within the community to develop powerful Large Multimodal Models (LMMs) featuring capabilities to conduct tasks across different modalities [72, 448]. For a better understanding of the historical development of LMMs based on NTP, we demonstrate a timeline in Figure 1, categorized by models' understanding or generation ability and different modalities.

In Figure 2, we use the image modality as an example to illustrate the workflow of **Multimodal Learning with NTP (MMNTP)**. The process can be divided into three key components: Tokenization, Modeling, and Training Objectives, which will be explained and discussed in details in the rest of the survey. For vision modality, image and video understanding capabilities have been demonstrated in large vision-language models such as GPT4-V [301], QwenVL [12], LLaVA [254], phi 3.5-Vision [1] and Gemini [370], while Emu [363] and Chameleon [369] show visual generation could be achieved in NTP manner. Similarly, end-to-end audio understanding and generation have been achieved in NTP-based models such as GPT4-o and Moshi [105, 302].

To equip LLMs with visual understanding capabilities, pioneering research such as Flamingo [3], BLIP2 [227], GPT4V [301], MiniGPT4 [509] and LLaVA [254] has demonstrated that LLMs can be easily adapted to process multimodal inputs such as images and videos, by converting multimodal information into tokens with a straightforward tokenization module, such as a visual encoder like CLIP [321] or a simple linear projection [18]. Subsequently, these models perform multimodal instruction tuning based on image-query-answer triples using the same NTP objective.

As Large Language Models bridge understanding and generation tasks in natural language processing, there is considerable interest in extending their capabilities to generate multimodal outputs. Recent advances in this direction include GPT-4o [302], which can understand and generate text, audio, and images using a unified multimodal LLM. We have also witnessed tremendous improvements from the open source community. For visual modality, Chameleon [369] and Emu3 [400] are two distinctive multimodal that unify understanding and generation in both language and image modalities. For audio, Moshi [105] can conduct tasks such as automatic speech recognition (ASR) and speech generation in an NTP manner based a pretrained LLM. As a general and fundamental approach, NTP also has promising implications for diverse fields like AI-for-Science such as designing proteins in biology [20] and composing molecule structure in chemistry [118].

To generate multimodal content using the NTP approach, it is crucial to recognize that unlike language, which is structured from discrete symbols, multimodal data like images and sounds inherently exist in a continuous space. A common technique to address this challenge is quantization. Vector Quantization (VQ) is a classical method that allows for the modeling of probability density functions for continuous multimodal data through discrete vector distributions [138, 306]. This technique aligns well with NTP modeling. With the rise of deep learning, neural VQ methods such as VQVAE [385] and VQGAN [112] have been developed, establishing a foundation for linking visual and audio generation with NTP. Significant work has emerged leveraging these VQ methodologies and the language modeling task. Examples include innovative systems such as DALL-E [327], CogView [91], CM3Leon [458], Parti [452], Muse [43], VideoPoet [199], LVM [13], Chameleon [369] and Infinity [149]. These methods often rely on external models, like VQGAN decoders, for image generation, making their approach a form of indirect multimodal generation. Parallel explorations have been conducted utilizing the NTP objective to directly generate images in continuous spaces, such as VAE's latent space [381], or by simulating a diffusion process [233, 507]. Unlike the indirect methods, only a few initiatives like ImageGPT [55] perform direct multimodal generation by predicting pixels from scratch. Additionally, NTP models can be augmented with various external models to facilitate multimodal generation. Notable examples include Emu [362], MiniGPT5 [506], and CoDi2 [368]. These approaches utilize the NTP framework to incorporate external diffusion models for image generation, showcasing another form of indirect multimodal generation.

We have covered powerful models that can understand or generate information across different modalities within the NTP paradigm. However, developing a single model that can both comprehend and produce information across multiple modalities, similar to human abilities, remains an intriguing goal in the pursuit of Artificial General Intelligence (AGI). Recently, a new research trend has emerged, focusing on the development of LMMs that unifies multimodal understanding and generation in the NTP paradigm. Notable examples include Unified-IO [269, 270], Chameleon [369], Transfusion [507], Show-o [422], Moshi [106], and Emu3 [401]. Unifying understanding and generation presents unique challenges, including the diversity of modalities and resolving conflicts between them. We will discuss these issues further in Section 6.

1.1 Overall Structure of the Survey

The structure of the survey is shown in Figure 3. Section 2 focuses on Multimodal Tokenization, and highlights the importance of tokenization as the bridge between raw multimodal data and their representations, distinguishing between discrete tokens that use Vector Quantization and continuous tokens. Section 3 delves into the Multimodal Backbone Model for NTP, indicating that an auto-regressive model, often resembling a large language model, is employed to capture multimodal tokens, utilizing distinct attention masks for different modalities to account for their specific features. Section 4 covers Training with Unified Multimodal Task Representation, explaining the training objectives varying from discrete to continuous token prediction, enabling multimodal

Table 1. Key Tables and Figures of the Survey.

Content	Section	Reference	Examples/Key-Words
	Tables		
Multimodal Tokenizers	§2 Multimodal Tokenization	Table 2	VQVAE [335], CLIP [321], HuBERT [160]
MMNTP Models	§3 Backbone Model for Multimodal Next Token Prediction	Table 3	Flamingo [3],DALLE [328], Unified-IO [271]
Multimodal Prompt Engineering	§4 Training with Unified Multimodal Task Representation	Table 4	Multimodal ICL, Multimodal CoT
Training Dataset	§5 Datasets and Evaluation	Table 5, 6	mC4 [436], Laion-5B [344],LLaVA [254]
Evaluation Dataset	§5 Datasets and Evaluation	Table 7	MME [120],MMBench [261],MMMU [466]
	Figures		
Historical Development of MMNTP Models	§1 Introduction	Fig. 1	Multimodal Generation and Understanding
Pipeline of MMNTP	§1 Introduction	Fig. 2	Tokenization, Modeling, Training Objectives
Illustration of Multimodal tokenizations	§2 Multimodal Tokenization	Fig. 4	Discrete Tokens, Continous Tokens
Tokenizer Training Methods	§2 Multimodal Tokenization	Fig. 5	Auto-Encoding, Contrastive Learning, ...
Illustration of Discrete Tokenization (VQ)	§2 Multimodal Tokenization	Fig. 6	VQVAE, Quantization, Codebook
Illustration of Continuous Tokenization	§2 Multimodal Tokenization	Fig. 7	Encoder, Decoder, Aligner
Categorization of MMNTP Models	§3 Backbone Model for Multimodal Next Token Prediction	Fig. 8	Compositional Model, Unified Model
Backbone of MMNTP Models	§3 Backbone Model for Multimodal Next Token Prediction	Fig. 9	Multimodal Transformer
Attention Patterns of MMNTP Models	§3 Backbone Model for Multimodal Next Token Prediction	Fig. 10	Causal, Semi-Causal, Prefix ...
Unified Structures for Vision Tasks	§3 Backbone Model for Multimodal Next Token Prediction	Fig. 11	VQA, Text-to-Image, Image-to-Image, ...
Unified Structures for Audio Tasks	§3 Backbone Model for Multimodal Next Token Prediction	Fig. 12	Audio Understanding, Audio Generation, ...
Training Objectives Explain	§4 Training with Unified Multimodal Task Representation	Fig. 13	Discrete/Continuous Token Prediction
Training Stage Overview	§4 Training with Unified Multimodal Task Representation	Fig. 14	Alignment, Instruction, Preference
Examples of Multimodal Prompt Engineering	§4 Training with Unified Multimodal Task Representation	Fig. 15	Flamingo [118], PCA-Bench [53]
Explanation of Multimodal Prompt Engineering	§4 Training with Unified Multimodal Task Representation	Fig. 16	ICL, CoT
Performance Comparisons on Understanding Tasks	§5 Datasets and Evaluation	Fig. 17	VQAv2 [136], MMMU [466]
Performance Comparisons on Generation Tasks	§5 Datasets and Evaluation	Fig. 18	Imagenet [342], GenEval [131]

output through VQ decoders or directly generating conditions for models like diffusion or VAE. The section also covers prompt engineering techniques such as In-Context Learning and Chain-of-Thought reasoning of MMNTP models adopted from LLM research. Section 5 introduces datasets and evaluation metrics, noting the superior performance of NTP models over non-NTP models in both understanding and generation tasks. Lastly, Section 6 outlines unsolved challenges in MMNTP research, such as scaling up MMNTP, emergent abilities, modality-specific biases, modalities interference, and MMNTP as universal interfaces, and discusses approaches to mitigate these challenges. Table 1 outlines key tables and figures in our survey.

1.2 Related Work

Several recent works have reviewed Large Multimodal Models (LMMs) in multimodal learning. For instance, Yin et al. [448] delve into the understanding capabilities of early vision-language models. Similarly, Awais et al. [8], Bordes et al. [27], Ghosh et al. [130], Caffagni et al. [36], and Zhang et al. [477] take a step forward and explore recent progress in multimodal learning with a focus on model architecture, training strategies, datasets, evaluation metrics and more. In addition, several surveys have reviewed multimodal learning in vision-language tasks, including pre-training [36], transfer learning [479], reasoning [402], and reinforcement learning from human feedback (RLHF) [479]. Beyond the discussions on the general revolution of the LMMs, specialized surveys have investigated the application of LMMs in domains such as multimodal agents [224, 421] and autonomous driving [72]. Recent surveys have also tackled key issues in multimodal learning, such as hallucinations in LMMs [256, 334] and efficiency of LMMs [185, 429].

Diverging from prior work that primarily focused on the understanding abilities of multimodal LLMs, our survey adopts a systematic perspective by integrating both understanding and generation in multimodal learning through the paradigm of next-token prediction. To the best of our knowledge, this is the first survey that reviews LMMs from the perspective of next token prediction, aiming to aid researchers in their exploration of multimodal intelligence.

In summary, in this survey, we aim to provide a holistic review on current multimodal models that rely on next token prediction. An associated GitHub link collecting the latest papers is at https://github.com/LMM101/Awesome-Multimodal-Next-Token-Prediction.

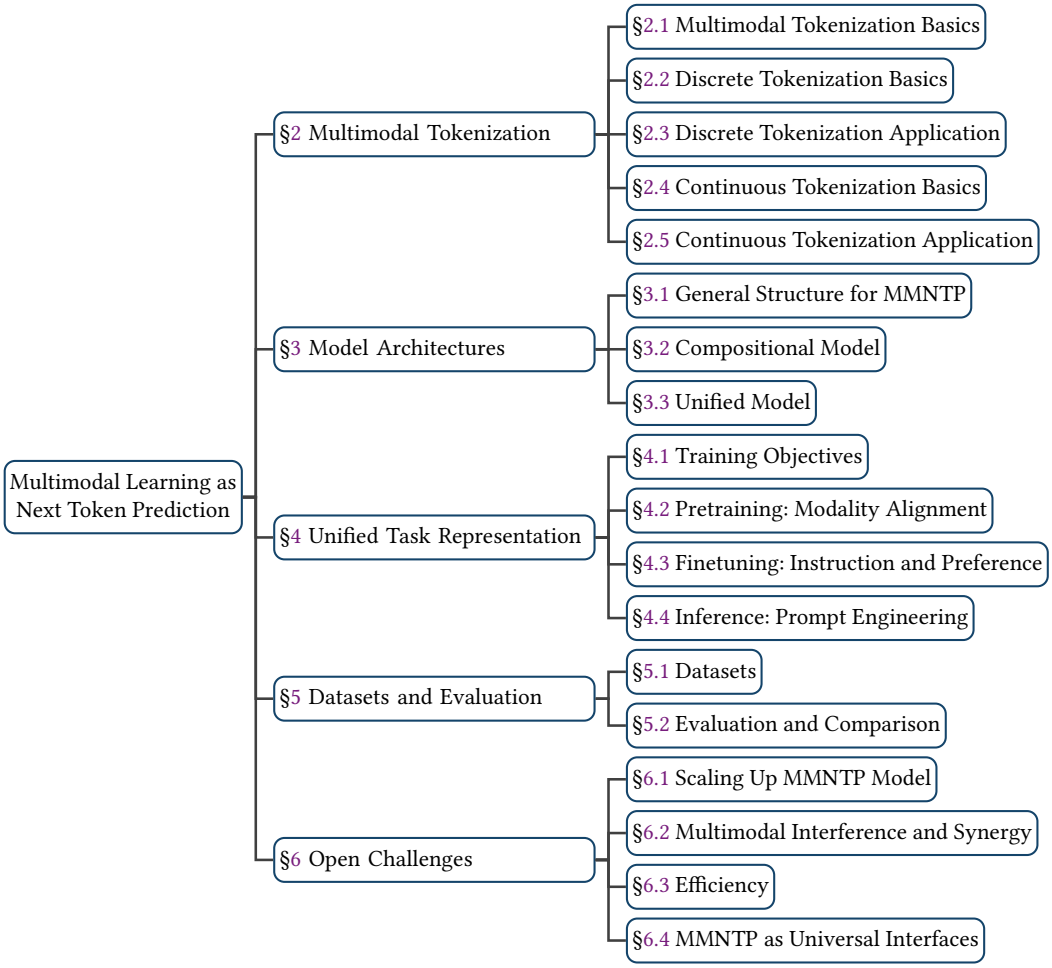

Fig. 3. Structure of the survey for Multimodal Learning with Next Token Prediction (MMNTP).

2 MULTIMODAL TOKENIZATION

Tokenization is the first and a fundamental step for multimodal sequential modeling under the next token prediction framework. It decomposes information from various sources, such as images, videos, and audio clips, into a sequence of minimal, manageable units known as tokens for the NTP model to learn. Table 2 provides an overview of the tokenizers used across various modalities in recent research.

Despite being derived from various modalities, these tokenization methods can all be categorized into two prototypes: **discrete tokenization** and **continuous tokenization**. In this section, we will initially introduce the general definition and basics techniques of training multimodal tokenizers (§ 2.1), then the fundamentals and applications of discrete tokens (§ 2.2, 2.3) and continuous tokens (§ 2.4, 2.5) in NTP framework.

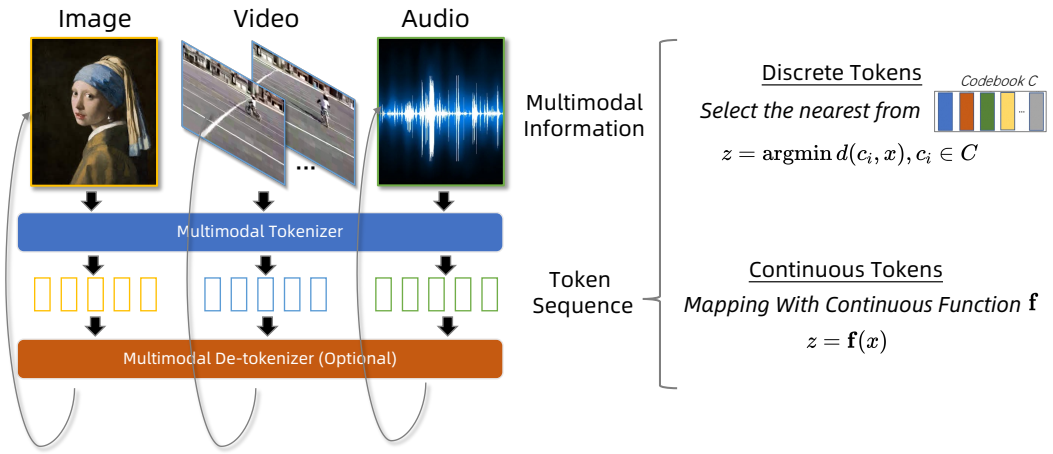

Fig. 4. Illustrations of multimodal tokenization.

2.1 Tokenization of Different Modalities

We first define the tokenization process as a function f that maps a sample x from the raw multimodal space X to a representation z in the tokenizer's output representation space Z_f.

$$f(x) = z, \tag{1}$$

where $x \in X$ and $z \in Z_f$.

2.1.1 Tokenizer Type. As illustrated in Fig. 4, tokenizers for multimodal information can be categorized into two types: discrete and continuous. This classification is based on how tokens are derived from the original data. Both tokenization methods encode the original information into a latent representation space, but they differ in their approach.

Discrete tokenization performs quantization on the latent space, utilizing a fixed-size, discrete space similar to the vocabulary of language models. In contrast, continuous tokenization does not involve quantization, resulting in a much larger representation space.

Discrete. In Equation 1, a discrete token implies that the representation space Z_f comprises a finite number of discrete symbols. The output space is called the codebook $C = \{c_1, c_2, ..., c_N\}$, where $c_i \in \mathbb{R}^0$, and each representation z is composed of codes from this codebook, i.e., $z = \{z_1, z_2, ..., z_n\}$ with $z_i \in C$. Language tokens are inherently discrete because they originate from a finite vocabulary. Each word or subword unit is mapped to a unique token from this predefined set. In contrast, modalities such as audio and images exist in continuous, high-dimensional spaces. To process these modalities within the same framework (i.e., NTP) as for discrete language tokens, they need to be transformed into a discrete representation.

Quantization is a process that maps values from a continuous space to a discrete space, typically resulting in a much smaller representation space. It is a default operation when a discrete representation is desired for tokenizing multimodal information. Quantization is often combined with auto-encoder techniques to reduce the size of the latent space. Typical examples include VQ-series tokenizers such as VQVAE [138] and VQGAN [112], which inherently feature discrete representations. Details of the quantization process are introduced in Section 2.2.

Table 2. Summary of tokenizers for different modalities. Training method includes AE (Auto-Encoding), DAE (Denoising Auto-Encoding), SP (Supervised Pretraining) and CL (Contrastive Learning), † means training with auxiliary losses.

Tokenizer	Year	Modality	Tokenizer Type	Backbone Structure	Training Method	Quantization Method	Used for Generation
VQVAE [335]	2017	Image, Video, Audio	Discrete	CNN	AE	Vanilla VQ	✓
3D ConvNets [89]	2017	Video	Continuous	3D-CNN	SP	✗	✗
VQVAE-2 [331]	2019	Image	Discrete	CNN	AE	Multi-scale VQ	✓
vq-wav2vec [10]	2019	Audio	Discrete	Transformer	CL	✗	✗
wav2vec 2.0 [11]	2020	Audio	Continuous	Transformer	CL	✗	✗
VQGAN [112]	2020	Image	Discrete	CNN	AE	Vanilla VQ †	✓
SoundStream [467]	2021	Audio	Discrete	CNN	AE	Vanilla VQ	✓
WavLM [56]	2021	Audio	Continuous	Transformer	DAE	✗	✗
HuBERT [160]	2021	Audio	Continuous	Transformer	DAE	✗	✗
NFNet [30]	2021	Image	Continuous	CNN	SP	✗	✗
BEiT [17]	2021	Image	Continuous	Transformer	DAE	Vanilla VQ	✗
CLIP [321]	2021	Image	Continuous	ViT	CL	✗	✗
ViT-VQGAN [450]	2021	Image	Discrete	Transformer	AE	Vanilla VQ†	✓
ViViT [6]	2021	Video	Continuous	ViT	SP	✗	✗
data2vec [9]	2022	Audio, Image	Continuous	Transformer	DAE	✗	✗
Whisper [322]	2022	Audio	Continuous	Transformer	SP	✗	✗
CLAP [110]	2022	Audio	Continuous	Transformer	CL	✗	✗
Encodec [104]	2022	Audio	Discrete	CNN, LSTM	AE	Residual VQ	✓
FlexiViT [24]	2022	Image	Continuous	Transformer	SP	✗	✗
Pix2Struct [218]	2022	Image	Continuous	Transformer	SP	✗	✗
RQVAE [216]	2022	Image	Discrete	CNN	AE	Residual VQ†	✓
MAGViT [454]	2022	Video	Discrete	3D-CNN	AE	Vanilla VQ†	✓
C-ViViT [388]	2022	Video	Discrete	Transformer	AE	Vanilla VQ	✓
CoCa [451]	2022	Image	Continuous	Transformer	CL+SP	✗	✗
EVA-CLIP [360]	2023	Image	Continuous	Transformer	CL	✗	✗
SAM-CLIP [392]	2023	Image	Continuous	Transformer	CL+SP	✗	✗
NaViT [79]	2023	Image	Continuous	Transformer	CL+SP	✗	✗
InternVIT [61]	2023	Image	Continuous	Transformer	CL+SP	✗	✗
SEED-Tokenizer [127]	2023	Image	Discrete	ViT	AE+CL	Vanilla VQ	✓
USM [487]	2023	Audio	Continuous	Conformer	DAE	✗	✗
DAC [204]	2023	Audio	Discrete	CNN	AE	Improved Residual VQ	✓
LMCodec [177]	2023	Audio	Discrete	CNN	AE	Residual VQ†	✓
HiFiCodec [439]	2023	Audio	Discrete	CNN, LSTM	AE	Group VQ	✓
SpeechTokenizer [486]	2023	Audio	Discrete	CNN, LSTM	AE	Residual VQ†	✓
MAGViT-v2 [456]	2024	Image, Video	Discrete	3D-CNN	AE	LFQ†	✓
LaViT [187]	2024	Image	Discrete	ViT+U-Net	AE	Vanilla VQ†	✓
Video-LaVit [186]	2024	Video	Discrete	ViT+U-Net	AE	Vanilla VQ, MPEG-4	✓
SPAE [455]	2024	Image	Discrete	CNN	AE	Vanilla VQ	✓
FACodec [189]	2024	Audio	Discrete	Transformer	AE	Group VQ	✓
SemantiCodec [255]	2024	Audio	Discrete	AudioMAE (ViT)	AE	Vanilla VQ	✓
WavTokenizer [178]	2024	Audio	Discrete	CNN, LSTM	AE	Vanilla VQ	✓
Mimi [78]	2024	Audio	Discrete	Transformer	AE	Residual VQ†	✓
VAR [373]	2024	Image	Discrete	CNN	AE	Multi-scale VQ†	✓
QwenVL2-ViT [395]	2024	Image, Video	Continuous	Transformer	CL	✗	✗

Continuous. In contrast to discrete tokenization, continuous tokenization represents data using a continuous space where tokens are derived directly from the data's inherent properties without enforcing quantization into a predefined codebook. In this approach, the representation space Z_f is not limited to a finite set of predetermined codes; rather, it preserves the continuous nature of the data. Each token z is sampled from a continuous distribution, allowing for a more nuanced and flexible representation that can capture the subtleties of the input data. Continuous tokenization is particularly advantageous for modalities that naturally exist in a continuous form and require a rich representational capacity to capture their complex patterns. For instance, in audio and visual data, continuous representations can effectively retain fine-grained temporal and spatial information that might be lost during discrete tokenization.

2.1.2 Features of Tokenizers. Before diving into different tokenization techniques, we summarize the basic two features (Representation and Reconstruction) that an ideal multimodal tokenizer should possess to achieve better understanding and generation capabilities in the NTP framework.

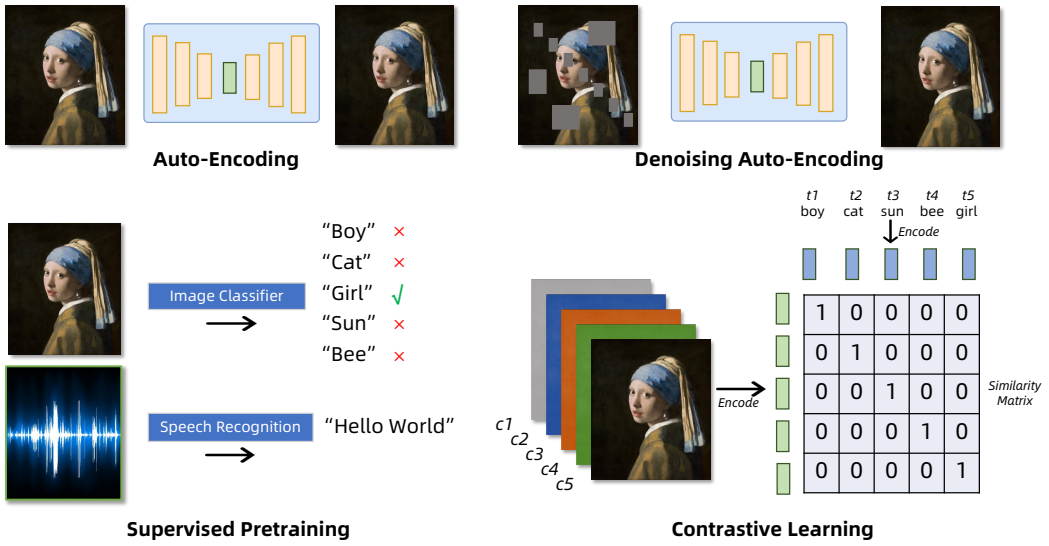

Fig. 5. Illustrations of the training method of tokenizers.

Representation Ability: Effective representation encodes semantically relevant information into the latent space Z while removing redundant information. This is crucial for various downstream tasks that learn a conditional probability $P(Y|X)$ over the label space Y, conditioned on the multimodal input space X, by replacing it with $P(Y|Z)$. Prominent tokenizers known for better representation include language-guided contrastive learning methods such as CLIP [321] and fully self-supervised methods like DINO [40].

Reconstruction Ability: For generating multimodal information, it is expected that the tokenization function f is invertible or nearly invertible, meaning there is a detokenization function g that can recover the original input from the representation space, satisfying $g(f(x)) = x$ or $g(f(x)) \approx x$. Notable works that excel in reconstruction include Auto-Encoder (AE) series models such as Variational Auto-Encoder [195] (VAE) and VQVAE [138].

It is important to note that these abilities are not mutually exclusive; their balance depends on the training techniques used.

2.1.3 Training Methods for Tokenizers. The training methodologies for tokenizers can be categorized into four groups, based on their respective training objectives: Auto-Encoding, Denoising Auto-Encoding, Supervised Training, and Contrastive Learning, as depicted in Figure 5. Herein, we provide a summary of the core concepts associated with various tokenizers.

Auto-Encoding. Auto-Encoder (AE) is a type of artificial neural network designed to learn efficient data representations. It consists of two main components: an encoder, which maps input data to a latent space with reduced dimensions, and a decoder, which reconstructs the input data from this latent representation. The training goal for an Auto-Encoder is to minimize the reconstruction error, ensuring the decoded output closely resembles the original input. Variants like Variational Auto-Encoders [195] (VAEs) use probabilistic approaches to generate more robust and informative embeddings. In multimodal generation models, tokenizers trained with auto-encoder methodologies are used to restore the multimodal input from the latent representation. A special case is diffusion models [90], which can also be viewed as an Auto-Encoder, enabling generation in a

non-autoregressive manner [233]. Discrete tokens are typically generated by quantizing [335] the continuous data representation within the latent space of auto-encoders.

Denoising Auto-Encoding. A Denoising Auto-Encoder (DAE) builds on the basic auto-encoder concept by introducing noise into the input data and training the model to reconstruct the original, noise-free version. This approach encourages the model to learn robust features capable of handling data corruption, thereby improving its generalization capabilities. In transformer-based models, a common technique known as Masked Language Modeling [84] involves masking parts of the input tokens and training the model to predict them, which can be viewed as a special type of denoising auto-encoder. This method has become mainstream across various modalities, popularized in language by BERT [84], in vision by BEiT [17] and MAE [153], and in audio by HuBERT [160].

Supervised Pretraining. Some tokenizers are pretrained on specific tasks using supervised learning, aiming to acquire task-specific representations through labeled datasets. These models are initially trained on large-scale datasets to capture specific features of the input data. In the vision modality, supervised tasks include semantic segmentation, object detection, and depth estimation. Models trained for these tasks, such as SAM [196, 392], ViTDet [240], and MiDaS [329], are later used in LMMs as tokenizers, like in DeepSeek-VL [268] and Cambrain-1 [376], to extract diverse visual features from input data. In the audio modality, Whisper [322] is trained with 680,000 hours of labeled audio data in a weakly supervised manner. Thanks to its robust and powerful speech feature extraction capabilities, Whisper is widely used in Speech LLMs [65, 162, 366] for extracting speech embeddings.

Contrastive Learning. Contrastive Learning is a self-supervised learning method that focuses on learning representations by distinguishing between positive and negative pairs. The core idea is to bring similar (positive) examples closer together in the representation space while pushing dissimilar (negative) examples further apart. The items in each pair can belong to the same or different modalities. For example, DINO [40] uses image-image pairs to enhance vision representation, while CLIP [321] employs text-image pairs to improve language alignment within vision representation.

Currently, LMMs that only feature multimodal understanding capabilities, such as InstructBLIP [74] and LLaVA [254], opt for tokenizers with superior representation abilities like CLIP [321], as they do not require reconstruction of the multimodal information. Conversely, LMMs supporting multimodal generation capabilities tend to choose VQVAE as the tokenizer, exemplified by models like Unified-IO [271], Chameleon [369], Emu3 [401], among others [128, 396, 405].

2.2 Discrete Tokenization Basics

Unlike the language modality, which inherently comprises discrete symbols (e.g., tokens or words), most other modalities naturally exist in a continuous space. To bridge the gap, the core technique is **Vector Quantization (VQ)**, which aims to map the original continuous information into a compressed, finite representation space, i.e. discrete tokens. The discrete tokens can have 2-dimensional or 3-dimensional structures for images and videos. These tokens are initially linearized based on a specific order, such as left to right and top to bottom, transforming them into a 1-dimensional sequence. This linearization allows for effective modeling usingthe next token prediction objective.

In this section, we will first elaborate on modern vector quantization techniques widely used as multimodal tokenizers, such as VQVAE (§ 2.2.1) and its variants. Following that, we will introduce the specific optimizations of discrete tokenization in different modalities (§ 2.3).

2.2.1 Vector Quantization Methods. The origins of VQ method trace back to the 1950s at Bell Laboratories, where researchers endeavored to optimize signal transmission through the development

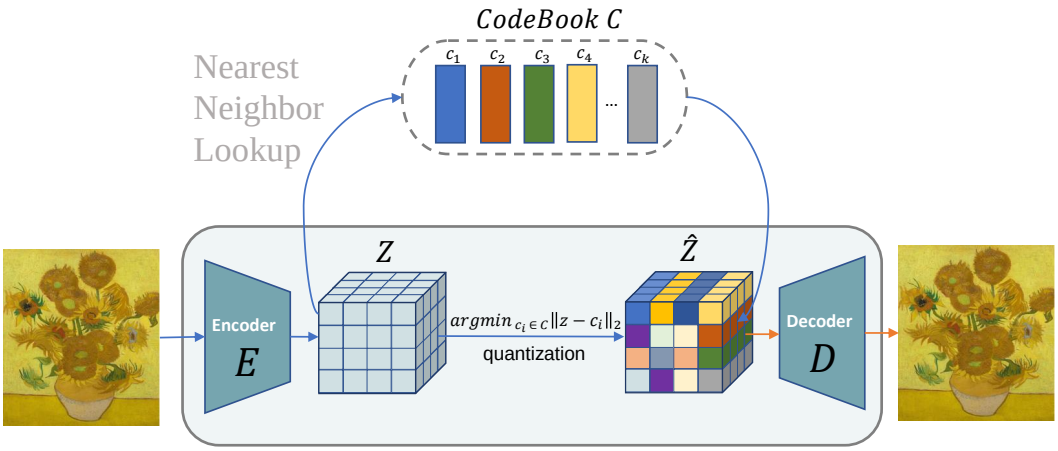

Fig. 6. Illustration of Vector Quantization. Blue lines denote the encoding and quantization process, while the orange lines denote the reconstruction process. The encoder transforms the input image into a latent representation, which is quantized by mapping each vector in Z to its nearest codebook entry in C. The quantized representation $\hat{Z}$ is then passed through the decoder to reconstruct the image.

of suitable discretization procedures [306]. In essence, quantization is the process of mapping an infinite set of continuous values to a smaller, discrete set of finite values. The primary objective of vector quantization is to reconstruct all the information in the original data as accurately as possible with a finite set of vectors, which is also called the *codebook*.

Vanilla VQ. The original VQVAE proposed by van den Oord et al. [385] is a milestone of many successive vector quantization methods. As shown in Figure 6, a VQVAE consists of three main components: the encoder, the quantizer, and the decoder. The encoder comprises the input data to a compact latent space, the quantizer select the nearest code vectors from the finite codebook to approximate the continuous latents, the decoder reconstruct the input data using the discrete codes. When training the VQVAE, three main loss components are crucial: reconstruction loss, codebook loss, and commitment loss [385]. The reconstruction loss, often implemented as mean squared error or binary cross-entropy, ensures accurate data reconstruction by minimizing differences between input and output. Codebook loss, or vector quantization loss, enables effective encoding by aligning encoder outputs with nearest codebook entries, ensuring discrete latent variables. Meanwhile, commitment loss acts as a regularizer, encouraging encoder outputs to stay close to codebook entries to maintain stable learning, preventing erratic mapping. As gradient can not pass the quantization operator (finding the nearest code), the straight-through estimator [21] is adopted to let the gradient flow normally.

Recent advancements in vector quantization methods have focused on achieving better image reconstruction and enhancing generative capabilities. To improve reconstruction quality, both architectural innovations and codebook designs have been proposed. Transformer-based frameworks, such as ViT-VQGAN [450], Swin-MAE [433], Swin-Unet [38], and Efficient-VQGAN [39], replace traditional CNN encoders and decoders with more robust modules like ViT [96] and Swin-Transformer [264, 265], leading to better feature representations and reconstruction fidelity. Additionally, several methods such as LFQ [456] and FSQ [291] are proposed to address the significant challenge of codebook collapse during **codebook learning**, where a large portion of code embeddings are not used when enlarging the codebook size, causing a redundancy in the

codebook and limiting the expressive power of the generative model [19]. For improved generative performance and efficiency, several approaches have been introduced. Tian et al. [374] propose Visual Autoregressive modeling, which facilitates image generation through "next-scale prediction", moving away from the traditional raster-scan "next-token prediction" used in standard VQVAE-based models. RQ-Transformer [216] employs residual quantization (RQ) to precisely approximate feature maps and reduce spatial resolution. RQ helps the RQ-Transformer to significantly reduce computational costs and effectively learn long-range interactions in inputs. RAR [459] introduces a randomness annealing strategy with a permuted objective, enhancing the model's ability to learn bidirectional contexts while retaining the autoregressive framework. TiTok [461] tokenizes images into 1D latent sequences, providing a more compact latent representation that is substantially more efficient and effective than conventional techniques. It greatly reduces the number of tokens required to encode an image compared to previous methods [39, 450].

VQ with Auxiliary Losses. The primary goal of the vanilla VQVAE is to accurately reconstruct input data by minimizing the mean squared error loss. However, this auto-encoding objective doesn't always align with human perception of the quality of reconstructed data. For example, in the visual modality, the vanilla MSE loss often results in images with blurred details, particularly in human faces [210]. To address this issue, several approaches introduce higher-level training objectives aimed at improving the overall quality of the output data. In the realm of vision, perceptual loss [188] is widely used to enhance the quality of reconstructed images by leveraging a pre-trained CNN. VQGAN [39] incorporates a discriminator network to enhance image fidelity by adding an adversarial training objective. The role of the discriminator is to discern between the reconstructed and original images, while the VQ-VAE is optimized to deceive the discriminator, thereby improving the quality of the reconstructed images. In the audio modality, it is essential to decouple the audio into its acoustic and semantic components to achieve both powerful audio reconstruction quality and LLM modeling. SpeechTokenizer [486] and Mimi [78] introduce the loss of semantic distillation at the first layer of Residual VQ, using self-supervised models, such as HuBERT [160] and WavLM [56].

Residual Vector Quantization. Residual vector quantization (RVQ) has been used for image [217] and audio [468] generation, where quantized codes are refined by storing additional quantized residuals. Lee et al. [216] propose the RQVAE that also introduces a residual quantization to recursively quantize the feature map in a coarse-to-fine manner, employing a fixed-size codebook to maintain both precision and code diversity.

Product Quantization. El-Nouby et al. [108] propose product quantization (PQ), to factor the codebook into a product of smaller codebooks, allowing for high-quality quantizers without the requirement of intractably large codebooks.

Multi-scale Quantization. Tian et al. [374] introduce the Visual Autoregressive modeling (VAR), which develops a multi-scale quantization autoencoder that encodes images into K multi-scale discrete token maps using a shared codebook. It aids the model in generating images through "next-scale prediction," instead of the raster-scan "next-token prediction" typically used in standard VQVAE-based models. The multi-scale quantization enables the model to learn visual distributions and demonstrates strong generalization capabilities.

Finite Scalar Quantization. To generate concise and expressive tokens using a larger token vocabulary and avoid codebook collapse, Mentzer et al. [291] propose finite scalar quantization (FSQ). FSQ projects the VAE representation down to a few dimensions that can be quantized into fixed values, creating an implicit codebook.

Look-up Free Quantization. LFQ [457] reduces the embedding dimension of the codebook to zero, effectively replacing the codebook with an integer set. It allows VQVAE to improve the quality of

image reconstruction and generation by vastly increasing the vocabulary size by magnitudes. For example, the rFID on Imagenet decreases from 2.5 to 1.4 when the LFQ vocabulary size increases from 2^10 to 2^16 on ImageNet dataset.

Embedding-Free Quantization. Maskbit [406] explores an embedding-free tokenization approach that utilizes binary quantization. It projects latent embeddings into K dimensions and then quantizes them based on their sign values to produce bit token representations. The generated bit tokens exhibit highly structured semantic representations, which are crucial for generation tasks.

Group Vector Quantization. Unlike RVQ which models the information residually, Group Vector Quantization models the information across different dimensions. In the audio domain, HiFi-Codec [439] proposes a group-residual vector quantization technique to reduce the number of codebooks, while FACodec [189] disentangles speech into prosody information, content information, and acoustic details using three-factorized vector quantizers.

2.2.2 Evaluation of VQ Tokenizers. When evaluating VQVAEs, two critical metrics are commonly considered: **reconstruction ability** and **generation ability**.

Reconstruction ability refers to how well the VQVAE can reproduce the original input data after encoding and decoding. This metric evaluates the fidelity of the model in terms of how accurately it can reconstruct the input data from its latent representations. L2 distance, Peak Signal-Noise Ratio (PSNR), and reconstruction Fréchet Inception Distance (rFID) are often applied to assess the reconstruction ability.

Generation ability assesses the model's capacity to generate new, plausible samples from the learned distribution in the codebook space. This metric evaluates the creativity and diversity of the VQVAE in producing new data that is consistent with the training data distribution. To quantitatively evaluate generation ability, metrics such as the Inception Score (IS) and generation Fréchet Inception Distance (gFID) [157] are often used.

rFIDs are often computed between ImageNet validation images and their reconstructed images. gFIDs are usually computed against the training set with ADM's evaluation suite [86].

2.3 Discrete Tokenization for Different Modalities

Generic quantization methods provide basic ways to convert continuous data into discrete tokens. However, there isn't a single quantizer that works well for all modalities because each modality has unique characteristics. Therefore, it is important to create specific tokenizers for each modality. This section will explain the unique features of different modalities and showcase some examples of tokenizers for images, audio, and video, among others.

2.3.1 Image. Images can be tokenized into discrete symbols with the previously introduced VQVAE structure. Compared to text tokens, images diverge in three fundamental aspects that significantly impact how they should be tokenized:

1. Rich Information Granularity: Unlike text, which primarily encapsulates high-level semantic meaning, images are contain with a myriad of perceptual details. These encompass low-level visual elements such as colors, shapes, and textures, alongside more abstract concepts like objects and actions.
2. Dense Information: Images inhabit a densely packed representational realm, where each pixel, across multiple dimensions including height, width, and color channels (RGB being a common example), carries information. This stands in stark contrast to the discreteness of text in nature, characterized by sequentially arranged words.

3. Two-Dimensional Spatial Structure: Images are inherently structured in two dimensions, spread across a grid defined by height and width. This 2D layout differs fundamentally from the straightforward, one-dimensional sequence that characterizes textual data, introducing unique complexities in their processing and analysis.

Given these differences, bridging the gap between text and image modalities in the training of LLMs based on discrete image tokens requires a robust image tokenizer, which must balance the fusion of sufficient alignment with LLM's language ability (referred to as "representation"), the retention of rich original image information (referred to as "reconstruction"), and the efficient use of tokens given the growing inference cost of transformer decoder (referred to as "token efficiency"). These factors possess a trade-off [127, 128, 357, 456], making it crucial for the construction of an image tokenizer to maintain equilibrium among these factors.

In terms of better representation, models like ViT [96] are commonly employed, often aligned with a text encoder through contrastive loss [312, 321], or aligned with text modalities through generative loss [451]. Additionally, modules like Q-Former [227] can also be used for image feature transformation [128, 227]. Consequently, the resultant image features integrate higher-level semantics and gradually compress high-dimensional images into lower-dimensional representations aligned with text. While the initial arrangement of image patches follows a raster order, preserving intrinsic sequential relationships, this configuration lacks causal semantics, posing challenges for language modeling.

Regarding reconstruction ability, an image decoder is often layered atop the image encoder to reconstruct the original image from its representation, incorporating reconstruction loss into the training process [112, 128, 187, 309]. Training labels typically use the original images, but with advancements in diffusion models, more research is incorporating latents for diffusion models as reconstruction labels [128, 187].

For token efficiency, modules like selectors or mergers for image tokens are utilized to truncate their length (i.e., the number of tokens per image). For instance, SEED-LLaMA [128] compresses longer image features encoded by ViT into 32 continuous tokens using a Causal Q-Former and then discretizes them through quantization. LaViT [187] first predicts whether each patch token should be selected using a shared MLP, and then compresses the image length by employing selected patches as queries and unselected patches as keys and values in cross-attention blocks [128].

Beyond these aspects, some studies also focus on the unique properties of specific image types or tasks. For example, VQ-IMG aims to enhance the modeling capabilities of image tokenizers for faces [124], while LVM integrates tasks like segmentation and object detection during the training of models based on VQGAN to enrich the representation of image tokens [13]. StrokeNVWA introduces a VQ-Stroke method to discretize vector graphic images into stroke tokens [367].

2.3.2 Audio. Raw audios are typically stored as 16-bit integer values with a sampling rate that exceeds tens of thousands values per second, which leads to extremely long sequences and renders next token prediction training more difficult. Versatile quantization methodologies have been investigated for audio tokenization. Initially aimed at audio compression, these methodologies have more recently been developed to create compact semantic and acoustic representations in the context of NTP language modeling.

As a traditional companding algorithm, μ-law/A-law algorithm is commonly employed in speech generative models such as WaveNet [383]. While this algorithm projects each audio frame to an 8-bit value, it does not reduce the sampling rate, thereby preserving overlong sequences. Self-supervised learned models have shown exceptional performance in various speech-related tasks, sparking interest in clustering their speech representations for speech quantization. The vq-wav2vec [10] uses either a Gumbel-Softmax or online k-means clustering to quantize the SSL-learned dense

representation. HuBERT [160] is trained with a masked prediction task, whose targets are obtained through k-means clustering of learned features from earlier iterations. Utilizing quantized tokens learned with Self-Supervised Learning (SSL), GSLM [207] and VQTTS [99] demonstrate faster speed in speech generation tasks compared with WaveNet. Because SSL tokens are extracted with highly abstracted semantics while discarding low-level acoustic information, the reconstruction quality is relatively low, and speaker identity is lost [28]. Neural codec models typically apply a VQ-VAE on the raw audios with residual vector quantization, exemplified by SoundStream [467] and EnCodec [104]. They are originally designed for audio compression, have the capability to encode waveforms into discrete codes and faithfully reconstruct them back into high-quality waveforms. Recently, they are widely used in audio generation models such as AudioLM [28], VALL-E [391] and their variants [148, 354, 397], and reach new state-of-the-art performance on various tasks. Compared with traditional μ-law/A-law algorithms, codec models can efficiently reduce the length of token sequences. It can also maintain multi-scale acoustic information indicating speaker identity compared with highly-abstracted SSL-learned discrete tokens such as HuBERT [160] tokens. Additionally, the codec models are typically off-the-shelf and lightweight.

Latest works have attempted to impose additional supervision on the discrete codes extracted by codec models. The objective is to enhance their ability to extract and encode higher-level semantic information, thereby improving language modeling. SpeechTokenizer [486] is an RVQ-based codec model, where its first-layer codebook incorporates semantic information through the semantic distillation process, using HuBERT [160] representations as the semantic teacher. Mimi, used by Moshi [105], further improves upon this by replacing the semantic teacher from HuBERT with WavLM [56]. Additionally, it isolates the first-layer codebook from the RVQ process to achieve better semantic and acoustic disentanglement. To enhance the compression rate, WavTokenizer [178] is capable of quantizing one-second audio into 75 or 40 tokens with a single quantizer.

2.3.3 Video. Compared to images, videos introduce an additional temporal dimension that must be considered during the tokenization process. A straightforward strategy is to utilize an image-based VQVAE model to tokenize the video frame-by-frame. This approach is employed by several multimodal foundation models, such as LVM [13], LWM [257], and Unified-IO series [270, 271]. However, a significant drawback of frame-by-frame tokenization is its inability to compress video data over time, resulting in a high degree of token redundancy across frames—particularly in long-form videos—thereby imposing substantial computational demands [353]. Furthermore, using an image-based tokenizer fails to model temporal relationships between frames, leading to issues of temporal inconsistency.

To address token redundancy and enhance temporal modeling, several studies have proposed training a 3D tokenizer that compresses videos across spatial and temporal dimensions. For example, VideoGPT [437] applies a 3D-CNN architecture in the encoder and decoder of the video tokenizer. C-ViViT [388] uses a transformer architecture to split videos into 3D cubes, which are then discretized into token IDs.

There are two additional desirable features for a video tokenizer: **(1) Joint Image-Video Tokenization**. The MAGVIT series [456] enables tokenizing images and videos with a shared vocabulary. To achieve this, the number of frames in an input video, T, must satisfy $T = 1 + n \times F_T$, meaning the video comprises an initial frame followed by n clips, each containing F_T frames. When $n = 0$, the video contains only the initial frame, thus simplifying the video to an image. Accordingly, both the initial frame and each subsequent clip are discretized into a $(1, H', W')$ token map, where H' and W' are the height and weight of the token map. **(2) Temporal Causality**. Compared to vanilla 3D architectures, using causal 3D architecture can ensure the tokenization and detokenization

of each clip depend only on the preceding clips, facilitating autoregressive modeling along the temporal dimension.

2.3.4 More Modalities. Modeling various information as discrete tokens has gone far beyond the traditional text, image, video and audio modalities. In the computer vision field, we can unify the output spaces of tasks like object detection, semantic segmentation, and depth mapping into images. These can then be tokenized into discrete image tokens, allowing us to train a single NTP model to handle all these tasks [13, 396, 399]. In **robotics and embodied AI** domain, the robots actions in response to the environments can be coded into various discrete tokens and learn the policy in NTP manner as shown in recent studies such as VIMA [183], RT2 [32] and Locomotion NTP [324]. In **AI4Science**, by factorizing various proteins into DNA token sequences, protein language models are capable of learning from a wide array of sequences that span the evolutionary tree. These models have demonstrated their efficacy as powerful tools for sequence design and protein engineering, as highlighted in studies [281, 340].

2.4 Continuous Tokenization Basics

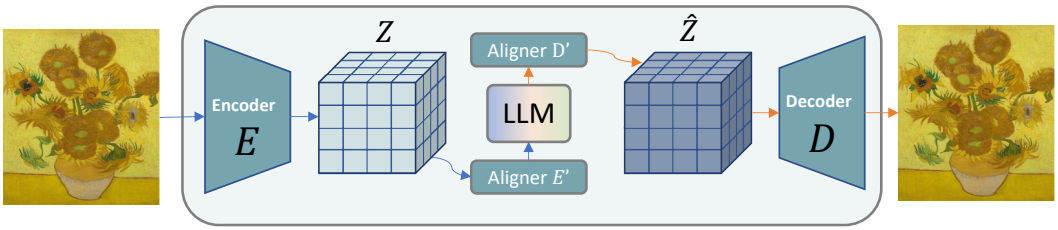

Fig. 7. Illustration of Continuous Tokens. Blue lines denote the encoding process, where the encoder transforms the input image into a latent representation Z, and aligner E' generates continuous tokens as input for the LLM to understand image content. Orange lines denote the generation process, where the LLM produces continuous tokens via aligner D', creating a latent representation $\hat{Z}$ for the decoder to reconstruct or generate an image.

Continuous tokens represent non-textual modalities in a continuous feature space, offering less information loss [51] and improved data representation compared to discrete tokens [423]. However, their dense feature encapsulation makes direct mapping to a fixed vocabulary challenging, unlike discrete tokens. It poses a challenge for LLMs aiming to comprehend and generate such information in a NTP manner.

To handle continuous multimodal token inputs for LLM to understand, transformations or adapters are necessary to balance data representation and text alignment. For multimodal generation, modifying the output head to align with non-textual modality specific decoders' input feature space is also crucial. The following subsections introduce the basic designs and change for LLMs to accommodate continuous multimodal token from multimodal understanding (§2.4.1) and generation (§2.4.2) perspectives.

2.4.1 Tokenize Continuous Input for Understanding. To effectively integrate raw non-textual modality data into Large Language Models (LLMs), two key steps are typically undertaken: (1) encoding the data into a more suitable representation space, and (2) aligning it with the LLM's feature space.

Encoding. The encoding of non-textual modality data aims to capture meaningful features and important nuances that are essential for the understanding of the data. This can be achieved through different types of encoders such as Transformer-based encoders [227, 253, 254, 321, 509]

or CNN-based encoders [3, 88, 183, 492]. There's also an option to go encoder-free [18, 194], which allows for raw data to be fed directly into the model.

Transformer-based encoders are widely used for their robust representation capabilities and generalizability [96, 386]. For a non-textual modality sample, the input is initially divided into patches and transformed into a 1D sequence, with each patch represented as a soft token. This sequence is then processed through the Transformer's encoder layers, employing self-attention mechanisms to capture relationships between patches. Consequently, the model produces a rich representation of the input. Typically, there are two types of encoders: (1) unimodal encoders, designed to process information from a single modality [6, 17, 96, 196, 242, 260, 265, 487]; and (2) multi-modal encoders, capable of integrating information from multiple modalities [110, 132, 294, 295, 321, 360, 451]. For instance, PaLM-E [97], Unified-IO-2 [269], and PaLI [58] use ViT [96] encoders trained solely on visual data. Conversely, LLaVA [254], Emu [357, 362], and Qwen-VL [12] utilize CLIP [321] or EVA-CLIP [360] encoders with contrastive loss to align textual and non-textual representations. NExT-GPT [417], CoDi-2 [368], and BuboGPT [502] employ ImageBind [132] as their non-textual encoder, aligning various modalities like audio, text, and heat maps with image representations.

In comparison, CNN-based encoders are less frequently used but remain vital due to their flexibility in image resolution generalization [454, 456] and ability to capture local features [183]. For example, DaVinCi [88] uses ResNet [154] as the visual encoder. Flamingo [3] utilizes NFNet [30], a normalizer-free ResNet, for image encoding.

Beyond encoders, Fuyu-8B [18] directly processes raw image patches after a single linear projection to accommodate images of varying resolutions and aspect ratios, similar to ViLT [194]. However, Fuyu-8B adds the flexibility of an any-resolution setting using a decoder-only model, benefiting from architectural simplicity but showing reduced downstream performance compared to encoder-based models. Moreover, ImageGPT [55] trains a decoder-only generative model on raw image pixel sequences, which, despite its effectiveness in image generation and understanding, requires significant computational resources and is limited to low-resolution images.

Input Alignment. After encoding non-textual modality data, we obtain a meaningful representation. However, this representation often lacks alignment with the textual embedding space of large language models, leading to a failure in properly understanding these inputs. Although multi-modal encoders like CLIP [321] have made strides in narrowing the gap, they still encounter two significant challenges: (1) the presence of redundant continuous tokens [3, 176, 227]; and (2) a lack of contextual semantics, such as causal semantics, because they are typically trained only with image-caption paired data rather than image-text interleaved data or image-prompt instructional data [127, 214, 370, 512]. Therefore, it is crucial to establish a connection between the representation space of non-textual modality data and the LLM textual embedding space. There are typically two approaches to construct such a bridge: (1) Slot-based Resampler [3, 227]; and (2) Projection [12, 18, 253, 254].

The Slot-based Resampler compresses redundant non-textual modality tokens from the encoding stage into fewer learned query vectors, known as slots. This is typically accomplished using multiple Transformer blocks with a cross-attention mechanism. For instance, BLIP-2 [227] employs a Q-Former and linear projection to bridge the image encoder with the LLM backbone. The Q-Former blocks consist of a self-attention layer on the learned queries, a cross-attention layer between the encoded image representation and the learned queries, and a feed-forward layer. Initially, it is trained for image-text matching, image-text contrastive learning, and image-grounded text generation, followed by training for next token prediction with the frozen LLM backbone. Another model using this approach is Flamingo [3], which utilizes a Perceiver Resampler [176] to compress

byte arrays into latent vectors in a modality-agnostic manner. Specifically, Perceiver [176] employs multiple cascaded attention mechanisms: the latents act as queries and initially cross-attend to keys and values calculated from the byte array (e.g., an image), followed by processing with a self-attention block, iterating several times. PerceiverIO [175] enhances this with an additional cross-attention block between an output query array and the slots (i.e., the latents). The Hierarchical Perceiver [41] decomposes the input array into multiple groups, compresses each group, and merges the resulting latents to obtain the output array.

Compared to a slot-based resampler, projection is much simpler in architecture, involving only a single linear projection [18, 254] or an Multi-layer Perceptron (MLP) [253]. For instance, LLaVA [254] employs a linear projection to convert encoded image representations into the language embedding space. Similarly, Fuyu-8B [18] projects raw image patches onto the embedding space. LLaVA-1.5 [253] enhances LLaVA by substituting the linear projection with an MLP.

There are also other approaches to connect the non-textual modality encoder with the LLM backbone. For example, Emu [362] leverages a Causal Transformer (i.e., C-Former) to convert the image tokens autoregressively; Emu2 [357] replaces the C-Former with mean pooling followed by a linear projection.

2.4.2 De-tokenize Continuous Output for Generation. The backbone of large language models is inherently designed for language generation. Typically, their output layers function as classification heads that predict distributions over a language vocabulary. For discrete non-textual modalities, the discrete token vocabularies can be integrated into the LLM's original text vocabulary since token generation is still managed by the classification heads. However, this approach does not work for continuous non-textual modalities. To enable the generation of continuous token outputs from LLM backbones, it is essential to modify their output layers (i.e., language modeling heads) to produce representations suited for non-textual modality data. These representations are then transformed to align with the input features of specific non-textual modality data decoders, such as a diffusion model [336]. Recent work includes MAR [234] and Transfusion [507]. We will further elaborate on the decoding of continuous output in §2.4.2 and the transformations to the output feature in §2.4.2.

Decoding. Unlike pure text generation, multimodal generation requires the model to decide when to switch modalities during decoding, due to their intrinsic differences. We refer to this objective as **positioning**. There are typically two methods to achieve this: (1) using placeholders [198, 417, 506]; and (2) employing a non-textual modality begin-of-sentence (BOS) token [95, 357, 362].

Firstly, special tokens can be introduced as placeholders for non-textual modality data. For instance, Mini-GPT5 [506] and GILL [198] utilize a sequence of image placeholder tokens ranging from [IMG1] to [IMGr], which can be interleaved with textual tokens, and these tokens are added to the model's vocabulary. Likewise, NExT-GPT [417] uses 5 image placeholder tokens, along with 9 audio and 25 video placeholder tokens. Secondly, the use of a single BOS token (sometimes accompanied by an EOS token) can simplify the process by signaling the position of non-textual modality data. For example, DreamLLM [95] employs a special <dream> token to mark the start of modality switching, allowing a single model run to process a sequence of queries. Emu [362] and Emu2 [357] use both image BOS and EOS tokens to encase encoded image features.

In addition to focusing on positioning, models must also learn to generate accurate features for non-textual modalities. Typically, the output layers of large language models (LLMs) feature classification heads for discrete token decoding, an objective we refer to as **output representation**. To enable continuous token outputs, modifications to these output layers are required. Generally, there are three approaches: (1) adapting the original language modeling head to be regressive [357,

362]; (2) introducing a new head for dense outputs [95]; and (3) utilizing the final hidden states before the language model head [198, 506].

Output Alignment. Typically, generated continuous tokens cannot be directly used for multimodal generation because they don't align with the input features of multimodal decoders like LDM [336] and AudioLDM [252]. To address this, additional modules are introduced to convert these tokens into representations suitable for multimodal decoders, ultimately generating the final non-textual modality data. For instance, NExT-GPT [417] employs a Transformer-based output projection, while Mini-GPT5 [506] and GILL [198] utilize a Q-Former-like architecture [227] consisting of a Transformer encoder and decoder to transform continuous tokens into conditional latent features for the Stable Diffusion Model. DreamLLM [95] uses a linear layer, whereas Emu [362] and Emu2 [357] directly utilize the generated continuous tokens as latents for multimodal decoders.

2.5 Continuous Tokenization for Different Modalities

While the aforementioned workflow and categorization outline a general approach to continuous multimodal tokenization, research indicates that employing modality-specific encoders, tailored to each modality, can significantly enhance performance [79, 295, 380]. Given the unique characteristics of different modalities, these approaches introduce specific inductive biases into the tokenization process.

2.5.1 Images. For images, specific research directions include but are not limited to: **image augmentation**, **resolution and aspect ratio** and **heterogeneous images**.

(1) Image Augmentation: This involves enhancing image representation using elements like depth, edge, and segmentation [196, 260, 392]. Prismer [260], for instance, introduces features beyond traditional RGB patches, such as depth and normal patchification. These features are compressed with a shared experts resampler before being integrated by a unified image encoder. SAM-CLIP [392] leverages SAM [196] and the CLIP text encoder for distillation training, boosting the semantic and spatial comprehension of the image encoder.

(2) Resolution and Aspect Ratio: This strategy includes support for high-resolution images, multi-resolution capabilities, and arbitrary aspect ratios [18, 79, 152, 430, 447]. For example, Fuyu [18] uses raw pixels as image encoding inputs for the LLM backbone via linear projection, employing a special image newline token for delineating raster-ordered patches. This enables support for various resolutions and aspect ratios. MS-ViT [152] suggests varying patchification based on image region complexity, introducing a gate mechanism to mark tokens needing finer patchification, which then undergoes encoding after position encoding interpolation.

(3) Heterogeneous Images: This includes encoding methods for specific image types like vector images, diagrams, charts, and PDFs [263, 431, 447]. Document images, for example, require detailed observation, as seen in TextMonkey [263], which splits large document images into smaller sub-images. Each sub-image is encoded individually, and trainable shifted attention layers are added post-frozen ViT layers for interactive representation across sub-images. These are then compressed and fed into the LLM backbone via an image and token resampler.

2.5.2 Audio. Recently, MELLE [289] indicates that predicting continuous tokens in an NTP manner can generate audio with high quality and naturalness comparable to ground truth. Traditionally, audio frames are converted from the temporal domain to the frequency domain using the Short-Time Fourier Transform (STFT) [139] or the Fast Fourier Transform (FFT) [103]. The magnitude of the Fourier-transformed frames is modeled as spectrogram, which is a 2D image showing how the frequency content of the signal evolves over time. Spectrograms or other transformations of raw audio signals are additionally going through the feature selection pipeline before converting

into discrete tokens. Mel-Frequency Cepstral Coefficients (MFCCs) [122] extracts coefficients that represent the short-term power spectrum of sound and is one of the most common features used in speech recognition. Mel-spectrogram [122] converts the spectrogram to the mel scale, which is more perceptually relevant to human hearing. These continuous features are commonly used in audio generation tasks.

Pre-trained foundation models, typically learned in a self-supervised manner on large-scale corpora, have emerged as powerful speech and audio representation extractors [211]. To obtain general speech features, wav2vec 2.0 [11] masks speech input in the latent space and addresses a contrastive task defined over quantized latent representations that are learned simultaneously. data2vec [9] biases the query-key attention scores with a penalty proportional to their distance. HuBERT [160] employs an offline clustering step to provide aligned target labels for a BERT-like prediction loss, which is applied solely on the masked regions. WavLM [56] introduces denoising in pretraining, jointly with regular masked speech prediction, as HuBERT. Whisper [322] is a speech recognition model characterized by an attention-based encoder-decoder architecture, trained on web-scale labeled speech data. It is increasingly being employed as a foundational speech model, extending its applications beyond speech recognition tasks [162, 287, 288, 366].

For continuous tokenization of audio, AST [134] uses a convolution-free pure-transformer architecture to extract features for audio classification, drawing insights from ViT [96]. Inspired by CLIP [321], CLAP [110] introduces a contrastive language-audio pre-training task to learn text-enhanced audio representations using supervised audio and text pairs. Fine-tuned based on a pre-trained CLIP model, Wav2CLIP [412] and AudioCLIP [146] incorporate an additional audio encoder using supervised pairs of audio and class labels. Audio-MAE [165] adopts a Transformer-based encoder-decoder framework to learn audio representations. Similar to MAE, it uses a reconstruction pre-training task where the decoder is tasked with reconstructing masked patches from the encoded information of the unmasked patches. BEATs [57] introduces a self-distilled tokenizer that converts continuous audio signals into discrete labels, facilitating classic mask and discrete label prediction pre-training.

2.5.3 Video. Video can be viewed as a sequence of images (frames) over time, making the modeling of temporal relationships between these frames a central focus. There are two common approaches to this modeling: post-temporal fusion and full-temporal fusion.

In the case of **post-temporal fusion**, models such as CLIP4Clip [276] and CLIPBERT [220] first independently encode each frame using an image encoder. They then employ lightweight pooling, convolution, and attention mechanisms to temporally fuse the features from all frames. The advantage of this approach lies in its ability to leverage pre-trained image encoders, thereby reducing the computational overhead associated with adapting to video data. However, a significant drawback is its limited capacity to adequately model features in the temporal dimension.

On the other hand, **full spatial-temporal fusion** models, like Temporal 3D ConvNets [89], VideoMAE [377], and ViViT [6], utilize 3D convolutions or 3D attention structures, allowing for comprehensive interaction among inputs in the spatio-temporal dimension. This enables better modeling of dynamic changes in temporal order, effectively capturing the motion of objects and backgrounds. However, this approach requires substantial 3D computation, prompting common strategies such as decoupling temporal and spatial self-attention [22, 332] and implementing sparse 3D attention [247] to enhance computational efficiency.

Recent advancements, such as TimeChat [333] and NumPro [419], have explored the integration of timestamp information into continuous video tokens, facilitating explicit time-vision associations for improved temporal grounding and reasoning.

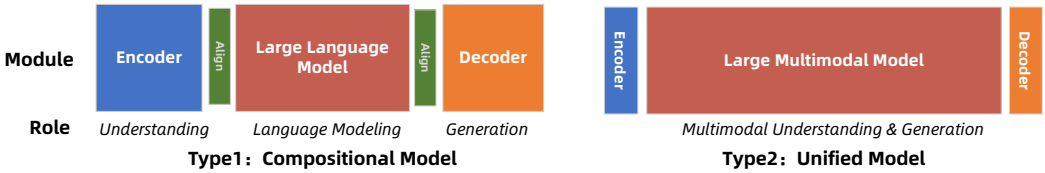

Fig. 8. Two types of Multimodal Next Token Prediction models. Compositional model utilizes powerful encoder and decoder for understanding and generation task, adding additional alignment layers. Unified model uses light-weighted encoder and decoder and leave most of the understanding and generation job to the backbone model.

Table 3. Summary of MMNTP model structures for different modalities.

Model	Year	Modality	Und.	Gen.	Task	Backbone	Tokenization	Encoder	Decoder
Flamingo [3]	2022	Image	✓	✗	I2T	Compositional	Continuous	NFNet [31]	✗
DALLE [328]	2022	Image	✗	✓	T2I	Unified	Discrete	dVAE [328]	dVAE [328]
Unified-IO [271]	2022	Image	✓	✓	I2T, T2I	Unified	Discrete	VQGAN [39]	VQGAN [39]
AudioLM [28]	2022	Audio	✗	✓	A2A	Compositional	Discrete	w2v-BERT [67], SoundStream [467]	SoundStream [467]
AudioGen [202]	2022	Audio	✗	✓	T2A	Unified	Discrete	Encodec [104]	Encodec [104]
MiniGPT4 [509]	2023	Image	✓	✗	I2T	Compositional	Continuous	EVA-CLIP [360]	✗
LLaVA [254]	2023	Image	✓	✗	I2T	Compositional	Continuous	CLIP [321]	✗
BLIP2 [227]	2023	Image	✓	✗	I2T	Compositional	Continuous	EVA-CLIP [360]	✗
Kosmos-1 [313]	2023	Image	✓	✗	I2T	Compositional	Continuous	CLIP [321]	✗
InternVL [61]	2023	Image	✓	✓	I2T	Compositional	Continuous	InternViT [61]	✗
QwenVL [12]	2023	Image	✓	✗	I2T	Compositional	Continuous	OpenCLIP [172]	✗
Fuyu [18]	2023	Image	✓	✗	I2T	Unified	Discrete	Image Patch [18]	✗
Molom [80]	2023	Image	✓	✗	I2T	Compositional	Continuous	CLIP [321]	✗
Codi-2 [368]	2023	Image, Video, Audio	✓	✓	I2T, T2I, A2T, T2A, T2V, V2T	Compositional	Continuous	ImageBind [132]	Diffusion(SD2.1) [336] + AudioLDM2 [258] + zeroscope[1]
MiniGPT5 [506]	2023	Image	✓	✓	I2T, T2I	Compositional	Continuous	EVA-CLIP [360]	Diffusion(SD2.1)[336]
Blip-Diffusion [226]	2023	Image	✗	✓	T2I	Compositional	Continuous	EVA-CLIP [360]	Diffusion(SD1.5) [337]
Kosmos-G [307]	2023	Image	✗	✓	T2I	Compositional	Continuous	CLIP [321]	Diffusion(SD1.5) [337]
Unified-IO2 [270]	2023	Image, Video, Audio	✓	✓	I2T, T2I, A2T, T2A, V2T	Compositional	Continuous+Discrete	AST [134] + OpenClip [172]	VQGAN [39]
Emu1 [362]	2023	Image	✓	✓	I2T, T2I	Compositional	Continuous	EVA-CLIP [360]	Diffusion(SD1.5) [337]
Emu2 [358]	2023	Image	✓	✓	I2T, T2I	Compositional	Continuous	EVA-CLIP [360]	Diffusion(SDXL) [316]
LaVIT	2023	Image	✓	✓	I2T, T2I	Compositional	Continuous	EVA-CLIP [360]	Diffusion(SD1.5) [337]
GSLM [208]	2021	Audio	✗	✓	A2A	Unified	Discrete	CPC [384] / HuBERT [160] / wav2vec [11]	Tacotron-2 [348]
SPEAR-TTS [193]	2023	Audio	✗	✓	T2A	Compositional	Discrete	w2v-BERT [67], SoundStream [467]	SoundStream [467]
Make-A-Voice [166]	2023	Audio	✗	✓	T2A	Compositional	Discrete	HuBERT [160], SoundStream [467]	SoundStream [467]
MusicGen [70]	2023	Audio	✗	✓	T2A	Unified	Discrete	Encodec [104]	Encodec [104]
VALL-E [391]	2023	Audio	✗	✓	T2A	Compositional	Discrete	Encodec [104]	Encodec [104]
SpeechGen [411]	2023	Audio	✗	✓	T2A	Unified	Discrete	Unit mBART [317]	Unit mBART [317]
MU-LLaMA [289]	2023	Audio	✓	✗	A2T	Compositional	Continuous	MERT [241]	✗
Pengi [82]	2023	Audio	✓	✗	A2T	Compositional	Continuous	CLAP [109]	✗
LTU [135]	2023	Audio	✓	✗	A2T	Compositional	Continuous	AST [135]	✗
SpeechLLaMA [413]	2023	Audio	✓	✗	A2T	Compositional	Continuous	Transformer	✗
SALMONN [366]	2023	Audio	✓	✗	A2T	Compositional	Continuous	Whisper [322], BEATs [57]	✗
Qwen-Audio [65]	2023	Audio	✓	✗	A2T	Compositional	Continuous	Whisper [322]	✗
AudioPaLM [339]	2023	Audio	✓	✓	A2T, T2A, A2A	Compositional	Discrete	w2v-BERT [67], USM [487], SoundStream [467]	SoundStream [467]
VIOLA [397]	2023	Audio	✓	✓	A2T, T2A, A2A, T2T	Compositional	Discrete	Encodec [104]	Encodec [104]
LauraGPT [102]	2023	Audio	✓	✓	A2T, T2A, A2A, T2T	Unified	Discrete	Encodec [104]	Codec Vocoder [102]
SpeechGPT [476]	2023	Audio	✓	✓	A2T, T2A, A2A, T2T	Unified	Discrete	HuBERT [160]	HiFi-GAN [200]
LlamaGen [356]	2024	Image	✗	✓	T2I	Unified	Discrete	VQGAN [39]	VQGAN [39]
VAR [373]	2024	Image	✗	✓	T2I	Unified	Discrete	Multi-scale VQVAE [373]	Multi-scale VQVAE [373]
DnD-Transformer[51]	2024	Image	✗	✓	T2I	Unified	Discrete	RQVAE [216]	RQVAE [216]
Mini-Genimi [245]	2024	Image	✓	✓	I2T, T2I	Compositional	Continuous	CLIP [321] + ConvNeXt [266]	SDXL[316]
Chameleon [369]	2024	Image	✓	✓	I2T, T2I	Unified	Discrete	VQ-SEG [124]	VQ-SEG [124]
MAR [234]	2024	Image	✗	✓	I2T	Unified	Discrete	✗	Mask-VQ
Fluid [113]	2024	Image	✗	✓	I2T	Unified	Discrete	✗	Random-VQ
Transfusion [507]	2024	Image	✓	✓	I2T, T2I	Unified	Discrete	CLIP [321]	AR-Diffusion [507]
VILA-U [420]	2024	Image, Video	✓	✓	I2T, T2I, T2V	Unified	Discrete	SigLIP [473], RQVAE [216]	RQVAE [216]
Show-o [422]	2024	Image, Video	✓	✓	I2T, T2I, T2V	Unified	Discrete	MAGVIT-v2	MAGVIT-v2
MIO [405]	2024	Image, Video, Audio	✓	✓	T2T, [I/V/A]2T, T2[I/V/A],[I/V]2[I/V], A2A	Unified	Discrete	Seed-Tokenizer [127], SpeechTokenizer [486]	Seed-Tokenizer [127], SpeechTokenizer [486]
Emu3 [401]	2024	Image, Video	✓	✓	I2T, T2I, T2V	Unified	Discrete	MoVQGAN [504][2]	MoVQGAN [504][2]
Janus [410]	2024	Image	✓	✓	I2T, T2I	Compositional	Continuous+Discrete	SigLIP [473]	VQGAN [356]
VoiceCraft [311]	2024	Audio	✗	✓	T2A	Unified	Discrete	Encodec [104]	Encodec [104]
BASE TTS [206]	2024	Audio	✗	✓	T2A	Unified	Discrete	WavLM [56], CNN	CNN
UniAudio [440]	2024	Audio	✗	✓	T2A, A2A	Unified	Discrete	Universal Codec [440]	Universal Codec [440]
CosyVoice [101]	2024	Audio	✗	✓	T2A	Compositional	Discrete	Conformer [141]	Transformer, ResNet
FireRedTTS [143]	2024	Audio	✗	✓	T2A	Compositional	Discrete	HuBERT [160], ECAPA-TDNN [83]	Transformer, ResNet
Seed-TTS [4]	2024	Audio	✗	✓	T2A	Compositional	Discrete	Unknown	Unknown
MELLE [289]	2024	Audio	✗	✓	T2A	Unified	Continuous	Mel-Spectrogram	Mel-Spectrogram
WavLLM [162]	2024	Audio	✓	✗	A2T	Compositional	Continuous	Whisper [322], WavLM [56]	✗
AudioFlamingo [201]	2024	Audio	✓	✗	A2T	Compositional	Continuous	Clapcap [111]	✗
SpeechVerse [76]	2024	Audio	✓	✗	A2T	Compositional	Continuous	WavLM [56], Best-RQ [64]	✗
VoxtLM [283]	2024	Audio	✓	✓	A2T, T2A, A2A, T2T	Unified	Discrete	HuBERT [160]	HiFi-GAN [200]
LLaMA-Omni [115]	2024	Audio	✓	✓	A2A	Compositional	Continuous	Whisper [322]	Transformer
Mini-Omni [424]	2024	Audio	✓	✓	A2A	Compositional	Continuous	Whisper [322]	Transformer
Moshi [78]	2024	Audio	✓	✓	A2A	Unified	Discrete	Mimi [78]	Mimi [78]
Qwen2VL [395]	2024	Image	✓	✗	I2T	Compositional	Continuous	Qwen2-VL ViT [395]	✗
EVE [87]	2024	Image	✓	✗	I2T	Unified	Discrete+Continuous	Image Patch [59] + CLIP [321]	✗
SOLO [59]	2024	Image	✓	✗	I2T	Unified	Discrete	Image Patches [59]	✗
MonoInternVL [275]	2024	Image	✓	✗	I2T	Unified	Discrete	Image Patch [275]	✗
RAR [459]	2024	Image	✗	✓	T2I	Unified	Discrete	MaskGit-VQGAN [44]	MaskGit-VQGAN [44]
Infinity [149]	2024	Image	✗	✓	T2I	Unified	Discrete	VAR [373]	VAR [373]

3 BACKBONE MODEL FOR MULTIMODAL NEXT TOKEN PREDICTION

After multimodal information is tokenized into sequential tokens, we need a model capable of handling multimodal information. In the literature, two classic MMNTP model structures are depicted in Fig. 8: 1) the Compositional Model and 2) the Unified Model. The key distinction lies in their design: the Compositional Model relies on heavily trained external encoders and decoders (such as [321]), and Diffusion models [158], for understanding and generation tasks

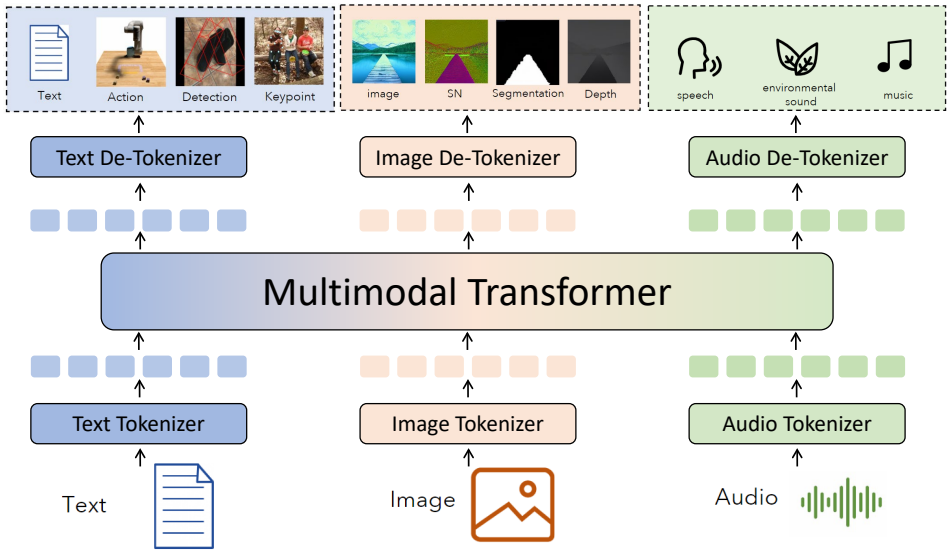

Fig. 9. Input text, images, audio, or image/audio history are encoded into sequences of tokens which are concatenated and used as input to an multi-modal transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip. Some image examples are referenced from [270].

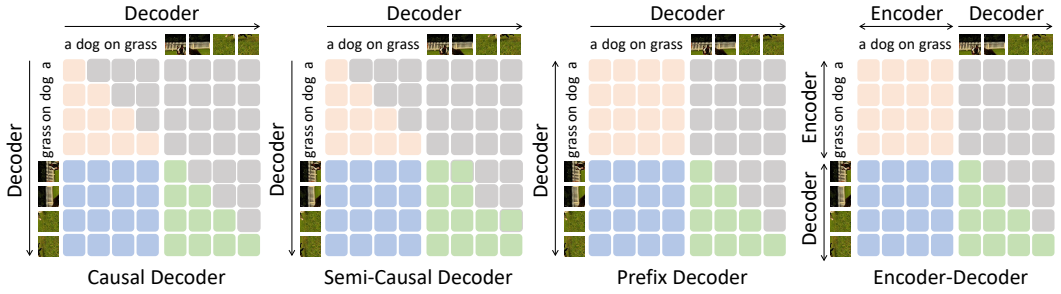

Fig. 10. Attention patterns in a causal decoder, non-causal decoder, and encoder-decoder architecture. In a causal decoder, each token attends to the previous tokens only. In a semi-causal decoder, each token could attends to part of future token and all past tokens [373]. In both prefix-causal decoder and encoder-decoder, attention is allowed to be bidirectional on any conditioning information. For the encoder-decoder, that conditioning is fed into the encoder part of the model.

respectively. In contrast, the Unified Model features lightweight encoders and decoders, with multimodal understanding and generation tasks primarily occurring within the backbone model, typically a large transformer decoder. A categorization of current MMNTP models is shown in Table 3. We will introduce the general structure of MMNTP model in Section 3.1, the recent advances in compostional and unified models in Sections 3.2 and 3.3, and compare them in Section 3.4.

3.1 Basic Structure of MMNTP Model

As shown in Fig. 9, to implement multimodal understanding and generation as next token prediction, this typically involves three steps. **Step 1.** Encode various inputs – images, text, audio, action,

boxes etc., into sequences of tokens in a shared representation space. **Step 2.** Use a multi-modal transformer to predict next token in an auto-regressive manner. **Step 3.** Decode the predicted tokens into the space of their respective modalities.

Fig. 9 also showcases the key modules of the NTP-based multimodal model, including tokenizers (encoders) and de-tokenizers (decoders) for each modality, as well as the multimodal Transformer. The tokenizer (encoder) and de-tokenizer (decoder) modules often appear together and are pre-trained using unimodal data through techniques such as reconstruction. They have the capability to split the original input into tokens using the tokenizer (encoder) and restore the tokens back to their original form using the de-tokenizer (decoder). Once all the tokenizers (encoders) and de-tokenizers (decoders) for each modality are pretrained, we can activate the required tokenizer (encoder) separately for tokenization of input containing multiple modalities, enabling us to obtain a multimodal token sequence. Finally, these multimodal token sequences are fed into the multimodal Transformer for NTP training.

For multimodal Transformers, we can use different attention masks to control the flow of information from different modalities [373, 422]. As shown in Fig 10, a common attention mask is the causal mask, which requires each token to only depend on preceding context for generation. However, certain tasks require generating subsequent text conditioned on a content-rich input prefix, such as generating summaries based on a rich-text-format document. For such tasks, we can also utilize a non-causal mask, which applies a bidirectional attention to the prefix, allowing the context within the prefix to interdepend and provide better representation, while using causal attention for autoregressive generation of the content to be generated. In summary, we can flexibly select attention masks based on the requirements of the task.

3.1.1 A Unified Structure for Vision Tasks. As illustrated in Fig. 11, various tasks in the vision modality can be encapsulated within the framework of MMNTP. Currently, a majority of large multimodal models (LMMs), such as LLaVA [254] and the Qwen-VL [12, 395] series, adhere to the NTP-based visual question answering paradigm. In this approach, images and text instructions are tokenized and sent to the transformer decoder to obtain the answer tokens. Another line of research, focusing on auto-regressive image generation, primarily adopts the NTP-based text-to-image generation paradigm, as seen in models like LlamaGen [356], VAR [373], and DnD-Transformer [51]. Alternatively, the output image tokens can be generated in a non-causal order, as demonstrated by works like MaskGIT [44] and RAR [459]. Additionally, these tokens can be continuous and later sent to a diffusion-based image de-tokenizer, as seen in recent developments like MAR [234] and Transfusion [507]. Some research combines the above paradigms to enable LMMs to perform both visual understanding and generation, as evidenced by models such as Show-o [422], Janus [410], and Emu3 [401]. Specifically, the NTP paradigm also supports various image-to-image tasks, such as image editing and semantic segmentation, as distinguished by Unified-IO2 and LVM [13].

3.1.2 A Unified Structure for Audio Tasks. As illustrated in Fig. 12, distinct NTP-based model architectures are required for various audio processing and generation tasks. For audio understanding [65, 162, 366], large-scale data pre-trained encoders demonstrate superior performance in extracting information from speech compared to discrete tokens. Additionally, an adapter is employed to facilitate the connection between the audio and text domains. Meanwhile, text instructions can specify specific audio processing tasks, such as automatic speech recognition, speech translation, and speech question answering. For audio generation, audio signals are typically transformed into discrete tokens [202, 206, 391] or continuous tokens [289]. These tokens can subsequently be converted back into waveform format through the use of corresponding decoders or vocoders. The text serves as either the specific speech content to be synthesized, or a detailed description of the audio. Leveraging the in-context learning capabilities and the scalability potential of the

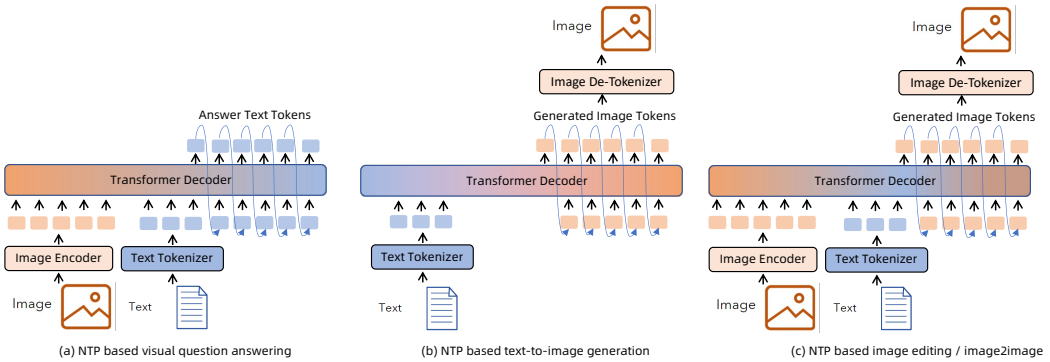

Fig. 11. Next token prediction based structures for (a) visual question answering, (b) text-to-image generation and (c) text guided image editing / image-to-image transform which require both image understanding and generation capabilities.

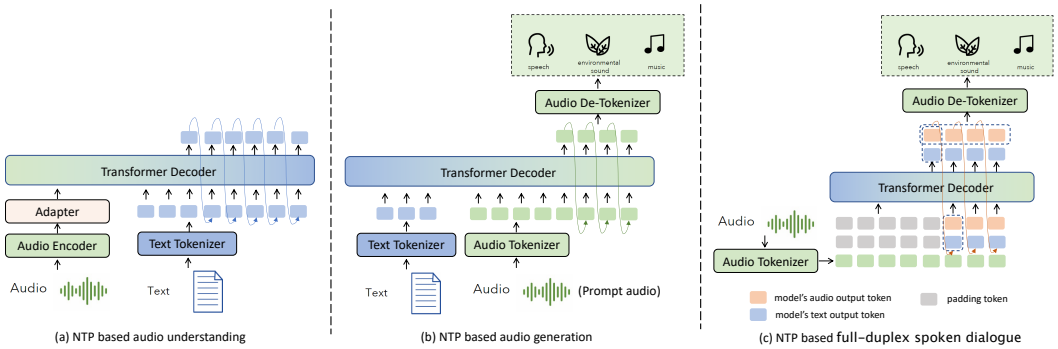

Fig. 12. Next token prediction based structures for (a) audio understanding, (b) audio generation and (c) full-duplex spoken dialogue, which requires both understanding and generation capabilities.

NTP based model, it achieves exceptional performance in zero-shot text-to-audio synthesis where a prompt audio is provided. Recently, the exploration of full-duplex real-time spoken dialogue [78, 302] has been progressing at a rapid pace, which requires strong audio understanding and streaming speech generation capabilities. In Moshi [78], to address these requirements, multiple audio streams, encompassing both user inputs and model outputs, are modeled concurrently, and a novel streaming audio tokenizer is introduced. For these tasks, the parameters of the Transformer decoder can be effectively initialized using those derived from an LLM.

3.2 Compositional Model

As shown in Fig. 8, the Compositional Model utilizes advanced external models to serve as the encoder and decoder for processing multimodal information. The section introduces the two components individually.

3.2.1 Connecting External Encoders for Understanding. A common architectural approach to enabling multimodal information understanding ability in LLM is using a robust external encoder to encode raw multimodal data to better representations. Pioneering work includes MiniGPT4 [509]

and LLaVA [254], which combine a vision encoder, an alignment layer and an LLM for general-purpose visual and language understanding. The LLaVA-style structure [254], which uses CLIP [321] as the encoder and an MLP as the alignment layer, has been utilized in numerous subsequent models. Recent studies reveal that scaling up the visual encoder [12, 60] and allowing for more flexible input image resolutions [12, 430] can significantly improve the model's visual perception abilities. Similar architectural approaches are employed within the audio domain to equip LLMs with the ability to perceive and process speech signals, as exemplified by models such as SALMONN [366], Qwen-Audio [65], and WavLLM [162]. For a detailed discussion on encoder design, please refer to Section 2.4.1.

3.2.2 Connecting External Decoders for Generation. To enable the LLM to generate multimodal outputs, including images, a straightforward approach is to connect it to a powerful image generation model such as a latent diffusion model [338]. In this context, it is crucial to ensure that the LLM generates continuous features beyond just language tokens, aligning the output with the input space of the diffusion models. Typical work includes Emu [362], which adds a regression head on top of the LLM's output hidden state to predict the visual embedding for the diffusion model. For a detailed discussion on decoder design, please refer to Section 2.4.2.

To enable both multimodal understanding and generation abilities of LLM in compositional manner, an external encoder and decoder can be attached to the backbone model simultaneously. A classic structure is exemplified by Emu1 and Emu2 [362, 363], which adopts EVA-CLIP [360] as the encoder and SDXL as the image decoder. For the audio domain, LLaMA-Omni [115] utilizes Whisper-large-v3 [322] as the encoder and a Transformer based decoder.

3.3 Unified Model

As shown in Fig. 8, the Unified Model leverages a light-weight encoder and decoder to process and generate multimodal information. The backbone model takes up most of the roles in understanding and generation tasks. This section will introduce two main structures of the unified model.

3.3.1 Quantization-based Autoregression. The quantization-based method is widely applied in building a unified model for multimodal understanding and generation due to its simplicity and similarity to the causal language modeling task. Typically, the encoder and decoder are derived from VQVAEs, trained to reconstruct the input from a discrete representation space. Focusing on generation, research explores generating images [51, 328, 356, 373] and audio [70, 202, 206, 440] with higher quality in an autoregressive manner and integrating advanced techniques for optimizing LLMs. Another line of work focuses on both understanding and generating multimodal information using quantization-based methods. Notable examples include Unified-IO [271], Chameleon [369], Emu-3 [400] and Moshi [78], which employ a unified NTP training objective for multimodal understanding and generation tasks.

3.3.2 Autoregressive Diffusion. The quantization-based method often faces criticism regarding generation quality. It typically produces images in a raster-scan order, which contradicts the intrinsic nature of 2D images. Additionally, the quantization process can lead to information loss. Several works aim to integrate the diffusion process into the NTP to enhance generation quality. Unlike compositional methods, the diffusion model is trained from scratch alongside the entire transformer model. Distinctive works such as Transfusion [507], MAR [234], CosyVoice [101] and Fluid [113] demonstrate that diffusion models can be jointly trained with language modeling tasks, offering superior image generation quality compared to quantization-based methods.

The debate between quantization-based and diffusion-based autoregressive methods for image generation is on-going, highlighting the need for further research. For instance, while many

diffusion-based AR methods [234, 507] claim better generation quality compared to quantization method, Emu3 [400] significantly outperforms diffusion baselines like SDXL using a quantization-based AR approach. DnD-Transformer [51] showcased that quantization-based AR generation has superior performance in generating rich-text images than diffusion models. In summary, it is not concluded yet which modeling method has superior performance than another currently.

3.4 Comparison Between Compositional and Unified Models

This subsection delves into a detailed comparison between compositional and unified models, evaluating their respective strengths and weaknesses in terms of general multimodal intelligence, training and deployment efficiency, and their potential to scale with increasing computational resources.

General Multimodal Intelligence. Unified models handle multimodal understanding and reasoning within a single backbone model, whereas compositional models assign different tasks to specialized external models. Although NTP has transformed language intelligence, its impact on multimodal intelligence remains uncertain. Given this context, unified models are closer to a multimodal foundation model [224, 401] due to its end-to-end nature and it may hold more potential than their compositional counterparts, as they rely on a single NTP training objective, making them easier to scale compared to multi-module systems. We will discuss the scaling behavior of MMNTP models in Section 6.1.

Training Efficiency. Compositional models benefit from leveraging highly specialized external encoders and decoders, often resulting in reduced training time for new tasks since these components are pretrained separately. This modular approach allows for targeted updates, reusing existing powerful models without the need for extensive retraining of the entire system. In contrast, unified models leave most of the understanding and generation responsibility to one backbone model, leading to sub-optimal performance given the same amount of computation [422]. This integrated training can be more resource-intensive, but it potentially facilitates a more coherent feature space across modalities within the LLM backbone, potentially enhancing overall performance on diverse multimodal tasks.

Deployment Efficiency. The unified model, particularly when using quantization-based methods, demonstrates significantly superior deployment efficiency compared to the compositional approach. A single unified transformer decoder backbone can effectively leverage the advanced techniques developed by the LLM community for accelerating both training and inference, such as Flash-Attention [75] and vLLM [205]. This capability is frequently cited as a key advantage of unified models, as highlighted by works like [356, 401].

4 TRAINING WITH UNIFIED MULTIMODAL TASK REPRESENTATION

Once content from various modalities has been tokenized into a sequence of tokens, with a unified backbone model, typically a decoder-only transformer model [386], we can undergo training to tackle a wide array of downstream understanding and generation tasks following different training objectives (refer to Section 4.1). The training tasks are primarily divided into two categories, which resemble the training of large language models: Pretraining (refer to Section 4.2) and Finetuning (refer to Section 4.3).

For a sequence of input tokens $x_{1\sim i-1} = \{x_1, x_2, \ldots, x_{i-1}\}$, the model predicts the next token $x_i \in V$. The general loss function f for a single prediction could be written as:

$$L(\theta) = f\left(y_i, p_\theta\left(x_i \mid x_{1\sim i-1}\right)\right), \tag{2}$$

where:

- $L(\theta)$ is the loss, parameterized by the model parameters θ and loss function f.
- V is the total vocabulary. We use V_T, V_M to denote text split and multimdoal split of the full vocabulary, V_S to denote the continuous tokens which are continuous vectors.
- y_i represents the target output for the next token. In supervised training, y_i is typically derived from labeled data, whereas in self-supervised training, y_i can be constructed from the data itself without explicit labels, often using the true next token from the input sequence. In special cases, y_i could involve multiple tokens, enabling parallel prediction of next tokens.
- f is cross-entropy loss when y_i is the discrete token distribution. f can also have different forms like mean-square error if y_i belongs to continuous tokens.

Different training tasks differ in the organization of given sequence $x_{1\sim i-1}$ and target label y_i. For self-supervised training, the sequence itself provides the target y_i, with the correct next token being used as the label. This allows the model to learn from the vast amounts of unlabeled multimodal data available, which consumes larger training resources. Supervised training would require explicit labeling of the next tokens, which can improve more specific downstream tasks at the cost of being more labor-intensive in the data collection period.

4.1 Training Objectives

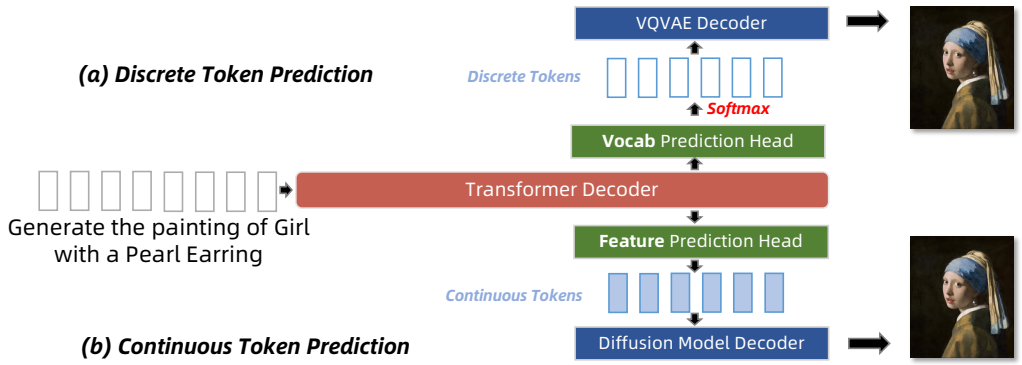

Fig. 13. Training objectives example for text to image generation. (a) For the discrete token prediction task, the backbone model, typically a transformer decoder, processes text input to generate discrete image tokens. This is achieved using a vocabulary prediction head and a Softmax function. The resulting tokens are then passed to the VQVAE decoder to construct the final image. (b) For the continuous token prediction task, there is no application of the Softmax function to map the output hidden state to a fixed-size vocabulary within the feature prediction head. Instead, the output is directly processed by a vision generative module, such as a diffusion model.

Based on what kind of target token y_i to predict, the NTP training objectives could be further categorized into two classes: **Discrete Token Prediction**, **Continuous Token Prediction** or a combination of them.

$$y_i \in \begin{cases} V_D, & \text{Discrete Token Prediction} \\ V_S, & \text{Continuous Token Prediction} \end{cases} \tag{3}$$

In Fig 13, we give an example using the task of text-to-image generation to show the difference between the two training objectives.

4.1.1 Discrete-Tokens Prediction (DTP). Discrete token Prediction (DTP) refers learn to predict the next discrete token given the context. The next token could belong to text or different modalities. This approach extends the conventional Causal Language Modeling (CLM), which typically deals with a unimodal text sequence, to accommodate inputs and outputs that interleave text with other data modalities, such as images. DTM enables the model to understand and generate different content from different modalities in a unified way. The training objective is to minimize the average cross-entropy loss among tokens.

Focusing on multimodal understanding ability, a majority of multimodal LLMs (e,g. Flamingo [282], GPT4V [301], MiniGPT4 [509], Qwen-VL [12] and LLaVA [254]) only predict language tokens V_T given multimodal inputs. It leverages the powerful reasoning ability and world knowledge of LLMs to support various multimodal understanding tasks without re-pretraining the model.

Enlarging the output token space to discrete multimodal tokens V_M like quantization codes would enable multimodal generation ability. In this approach, multimodal contents are first converted into discrete tokens, utilizing cross-entropy loss as the loss function. A major line of works is auto-regressive multimodal information generation, such as DALLE [327], CogView [91], Unified-IO [271], LVM [13] and Video-Poet [199].

Merging the two output spaces (V_T and V_M) into one model is an intriguing direction [257, 270, 271], which naturally unifies multimodal understanding and generation tasks. However, some related research [489] shows that learning to predict text tokens have no benifit for predicting multimdoal tokens and sometimes lead to strong conflict. Under the NTP training framework, whether multimodal generation helps understanding ability also remains unclear. Consequently, effectively integrating the output spaces of text and multimodal tokens presents itself as one of the main challenges in the domain, underscoring the need for innovative and scalable approaches to harness the full potential of NTP models in the realm of multimodal learning.

A variant of standard next token prediction is to predict multiple tokens at one time, disobeying the causal order. Recent researches [44, 374, 456] have found that parallel prediction is more effective for visual domains such as images and videos than simple raster-based prediction, which predicts the image tokens from left to right and top to down. MaskGIT [44] and MAGVIT [456] predict a portion of tokens at each prediction step according to a dynamic confidence threshold. VAR [374] predicts the visual tokens in a resolution-autoregressive manner, which predicts tokens in the same resolution in parallel and predict low-to-high images in sequential. Those approaches inject different inductive bias for different modality during NTP modeling, which is also an important challenge when unifying multiple modalities in multimodal NTP framework.

4.1.2 Continuous Token Prediction (CTP). In addition to discrete multimodal tokens, the multimodal information can also be represented as continuous vectors, referred to as Continuous-tokens. The Continuous-tokens can be viewed as conditions for external model such as stable diffusion model for better generation quality. The continuous tokens are usually predicted auto-regressively with MSE loss [198, 357, 362, 368, 506]. For example, Emu-1 and Emu-2 [357, 362] leverage a large language model to generate continuous tokens, which are used as condition for a pretrained diffusion model to generate images. The language model and diffusion model are trained simultaneously during the text-to-image instruction tuning stage. This method utilizes the powerful image generation ability of open-source diffusion model and unlocks the multimodal generation ability of large language model with modest additional cost.

Beyond utilizing continuous tokens as conditions for external models, some researches explored using continuous tokens to directly generate images, replacing discrete tokens with continuous tokens throughout the NTP training paradigm. El-Nouby et al. [107] reveals that when trained with L2 loss, a patch-based image Transformer exhibits scaling properties akin to those of LLMs. [237]

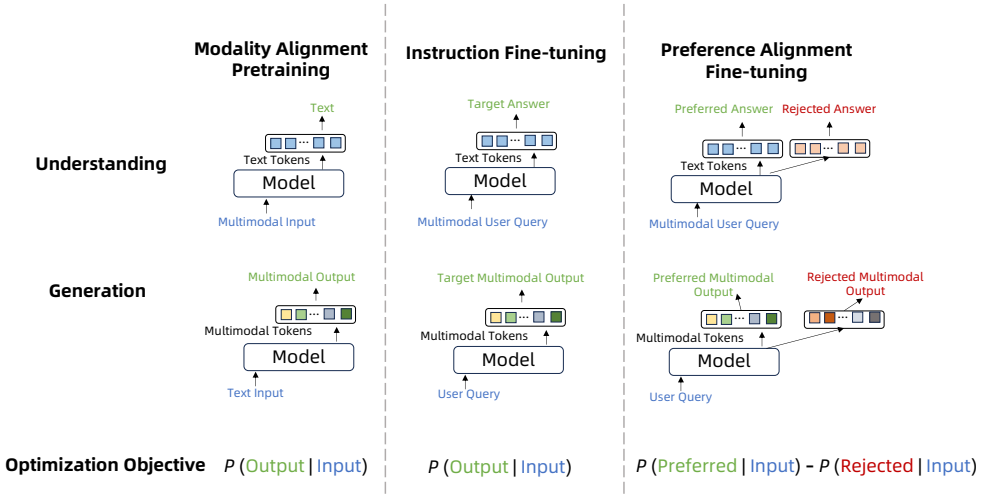

Fig. 14. Training stage overview.

represents image with continuous tokens and involves diffusion loss during training the causal transformer model. However, these models are trained solely on single modality such as image. Whether different training objectives for different modalities can coexist harmoniously in one NTP model remains under-explored.

4.2 Pretraining: Modality Alignment

Large Language Models have demonstrated their effectiveness and scalability in the pure language domain. In a similar vein, pioneering research is exploring the use of the abundant supply of multimodal data in training large multimodal models in NTP framework. The major focus of pretraining in LMM is to align the representation space of different modality with language space, which could be categorized into alignment in understanding (Section 4.2.1) and generation (Section 4.2.2) task.

4.2.1 Modality Alignment in Understanding. Modality alignment is a critical process that endeavors to represent inputs from diverse modalities within a shared space for subsequent processing. Given the inherent differences in the nature of various modalities, dedicated encoders tailored to each modality transform raw inputs into a vector representation, which is then aligned in the shared space. For instance, the alignment training of vision-language models typically occurs on a large-scale corpus $C = \{(C, I)\}$ comprising image-text pairs with the image denoted as I and its corresponding caption as C. The modality alignment objective typically adheres to a conditional language modeling format, expressed as:

$$L(\theta_{\mathcal{M}}) = f\left(y_i, p_\theta\left(x_i \mid x_{1\sim i-1}, I\right)\right), \tag{4}$$

where the parameter of the modality encoder module $\theta_{\mathcal{M}}$—such as a CLIP vision encoder responsible for mapping multi-modal inputs into vectors in the shared space—is exclusively trained to enhance stability.

It is noteworthy that the modality condition I for images can be seamlessly adapted to other modalities, such as videos and audios, with corresponding training corpora like WebVid [14] for video-text alignment and Clotho [98] for audio-text alignment, CroCo [409] for 3D views and embodiment Habitat [343]. Besides, it is also possible that the text and the image are interleaved

with each other, and the objective can be adjusted accordingly [7, 214]. We provide a comprehensive list of modality alignment training in the later section (§ 5.1.1).

4.2.2 Modality Alignment in Generation. The alignment objective can be easily adapted to the generative scenarios by replacing the one-hot word index y_i with corresponding modality tokens, which might be learned via a pre-defined codebook or optimized via regression. Take the traditional text-to-image task as an example, given a description-image pair (C, I), the alignment objective becomes:

$$L(\theta_{\mathcal{M}}) = f\left(y_i, p_\theta\left(t_i \mid t_{1\sim i-1}, C\right)\right). \tag{5}$$

In DTM, the y_i could be a targeted discrete visual token learned via an off-shelf model such as VQGAN, and the image content would be reconstructed by mapping the token back to the image space via the codebook. In CSM, the y_i is instead a contiguous modality vector that can be further decoded by a decoder to produce the image pixels [363]. Besides, the objective can also be implemented in a span corruption style for a better reconstruction of specific modalities [271].

Given that a primary objective in the alignment stage is to harmonize the semantics of concepts expressed across different modalities, comprehensive coverage of the training corpus becomes imperative. Consequently, the alignment training is often performed on web-scale datasets. For example, the visual-text alignment is usually conducted on up to millions and even billions of pairs on Laion400M [345] and Laion5B [344].

4.3 Finetuning: Instruction and Preference

After modality alignment training, LMMs acquire a foundational understanding of the semantics associated with various modalities in a unified semantic space. To further enhance LMMs' ability to comprehend and perform complex user queries, such as image understanding and generation, researchers employ *instruction tuning* on meticulously curated datasets. Subsequently, *preference alignment training* is utilized to refine model behaviors with implicit human preferences and address potential issues that may have emerged during earlier training phases. In the following discussion, we will discuss recent advancements in instruction tuning (§4.3.1 and §4.3.2) and alignment training (§4.3.3 and §4.3.4), as well as explore promising avenues for future research in these domains.

4.3.1 Instruction Tuning in Understanding. After the modality alignment training, different modality inputs now can be represented in a unified embedding space for the backbone LLM to perform complex tasks. Instruction tuning (*alias* supervised fine-tuning) plays a crucial role in activating this potential of multi-modal language models. Specifically, the instruction tuning aims to improve the model's ability to satisfy user queries. Again, take the vision language models as an example. The visual instruction tuning involves training the model on a dataset that usually consists of a multi-modal triplet (I, Q, A) of an image I, a user query Q, and a desired response A. This still can be achieved by the previous training object:

$$L(\theta) = f\left(A, p_\theta\left(x_i \mid x_{1\sim i-1}, I\right)\right). \tag{6}$$

Different from the previous alignment training, the instruction tuning stage involves a more challenging objective to reason over the modalities, motivating the model to explore the inner interaction between different modalities to increase the likelihood of the preferred answers. It has been shown that the quality of the instruction tuning is the key to the ability [253]. Pilot studies explore various methods for constructing high-quality instruction tuning datasets such as adapting publicly available multi-modal benchmarks [231, 432, 434], synthesizing datasets using self-instruction with ChatGPT/GPT-4 [50, 254, 496, 503]. Furthermore, mixing the multi-modal instruction dataset with text-only query-response pairs is also shown to be effective for improving

the instruction following ability [254, 434]. For A curated list of these instruction tuning dataset can also be found in later section.

4.3.2 Instruction Tuning in Generation. Similar to the practice in understanding, the key to improving the generation ability after alignment is collecting high-quality and diverse task datasets, where the reconstruction targets vary according to the task requirements. However, most training objectives still fall into the token modeling paradigm with different tokenization schemas. The desired output such as textual sentences, images/videos and audios, is represented in a sequence of N tokens $S = (s_0, \ldots, s_N)$, given the conditioned user queries Q specifying the requirements on the target outputs. During the instruction tuning stage, the following objective is optimized:

$$L(\theta) = f\left(y_i, p_\theta\left(s_i \mid s_{1\sim i-1}, Q\right)\right), \tag{7}$$

where y_i would be the corresponding discrete token or contiguous vector processed as in the alignment training objective. To provide wide coverage of the generation ability, previous work [271] ensembles a massive multi-tasking dataset and the sampling ratio during training would be balanced to better expose the model to underrepresented tasks. AnyGPT [474] utilizes commercial image-generation and music-generation systems to construct a large-scale high-quality text-to-multimodal instruction tuning datasets.

4.3.3 Preference Alignment Training in Understanding. Despite the progress made by previous training stages, misalignment issues that pose a potential risk of generating misleading content without anchoring to the provided visual context [238, 364], or biased responses against minority groups [301], still exist. To further align with human preference for LMMs, pilot studies draw insights from LLMs and apply alignment techniques such as Reinforcement Learning with Human Feedback (RLHF) [304] and Direct Preference Optimization (DPO) [325] for LMMs. LLaVA-RLHF [364] first explores the RLHF for VLM, by training a factuality-oriented reward model on a synthesized dataset to guide the VLM to produce outputs that anchor with the visual context better. Formally, let x be a prompt containing both images and text inputs, and y_i denotes the corresponding response generated by model π_i. The RLHF process can be formulated as:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x)\|\pi_{\mathrm{ref}}(y \mid x)\right],$$

where r is the reward model and the KL term penalizes deviations of the current model π_θ from the initial model π_{ref}. β is a hyper-parameter. The RLHF process aims to finetune the model to achieve higher rewards from the reward model, all while preserving the majority of its original knowledge.

As training the reward model can be difficult due to the stability issue, there has been a DPO method to tackle these challenges. The key insight behind DPO is that the optimal policy π^* has a closed-form solution with regard to a reward function r and initial policy π_{ref}:

$$r(x,y) = \beta\frac{\pi^*(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + \beta\log Z(x),$$

where Z is the partition function.

Under the Bradley-Terry (BT) preference model [29], the objective becomes:

$$\max_{\pi_\theta} \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)} - \beta\log\frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\mathrm{ref}}\left(y_l \mid x\right)}\right), \tag{8}$$

where σ denotes the sigmoid function. RLHF-V[463] collects human preference in the form of segment-level corrections on hallucinations, and performs dense direct preference optimization over the human feedback. Li et al. [230] build VLFeedback by annotating the preference with GPT-4V models and applies DPO on Qwen-VL-Chat showing clear advantages.

4.3.4 Preference Alignment Training in Generation. Due to the computation cost and the difficulty of collecting large-scale comparison datasets (i.e., creating slightly different images), there are few explorations on preference alignment in generative unified multimodal models. There are pilot studies investigating preference alignment for diffusion models, where the expected reward of a generated sequence $\boldsymbol{x}_{1:T}$ given a condition $\boldsymbol{c}$ and initial latent $\boldsymbol{x}_0$ is:

$$r\left(\boldsymbol{c}, \boldsymbol{x}_0\right) = \mathbb{E}_{p_\theta\left(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0, \boldsymbol{c}\right)}\left[R\left(\boldsymbol{c}, \boldsymbol{x}_{0:T}\right)\right] \tag{9}$$

Similar to the alignment training in understanding tasks, the objective is to maximize the expected reward while minimizing the KL divergence between the learned distribution p_θ and a reference distribution p_{ref}:

$$\begin{aligned} &\max_{p_\theta} \mathbb{E}_{\boldsymbol{c} \sim \mathcal{D}_c, \boldsymbol{x}_{0:T} \sim p_\theta\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right)}\left[r\left(\boldsymbol{c}, \boldsymbol{x}_0\right)\right] \\ &- \beta \mathbb{D}_{\mathrm{KL}}\left[p_\theta\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right) \| p_{\text{ref}}\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right)\right] \end{aligned} \tag{10}$$

Current methods for aligning image generative models mainly adopt DPO to bypass the cumbersome reward modeling process. Wallace et al. [389] re-formulate DPO to account for the intractable likelihood in diffusion models, where the evidence lower bound (ELBO) is employed to derive a differentiable objective function for optimization. The final DPO-Diffusion loss function encourages the model to improve the denoising process for preferred images more than for non-preferred images.

$$\begin{aligned} L_{\text{DPO-Diffusion}}\left(\theta\right) = -\mathbb{E}_{\left(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l\right) \sim \mathcal{D}, t \sim \mathcal{U}(0,T), \boldsymbol{x}_{t-1,t}^w \sim p_\theta\left(\boldsymbol{x}_{t-1,t}^w \mid \boldsymbol{x}_0^w\right), \boldsymbol{x}_{t-1,t}^l \sim p_\theta\left(\boldsymbol{x}_{t-1,t}^l \mid \boldsymbol{x}_0^l\right)} \\ \log \sigma\left(\beta T \log \frac{p_\theta\left(\boldsymbol{x}_{t-1}^w \mid \boldsymbol{x}_t^w\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{t-1}^w \mid \boldsymbol{x}_t^w\right)} - \beta T \log \frac{p_\theta\left(\boldsymbol{x}_{t-1}^l \mid \boldsymbol{x}_t^l\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{t-1}^l \mid \boldsymbol{x}_t^l\right)}\right), \end{aligned} \tag{11}$$

where condition $\mathbf{c}$ is omitted for brevity. The models are trained on the Pick-a-Pic [197] dataset, which contains pairwise preferences for images generated by SDXL-beta and Dreamlike, a fine-tuned version of Stable Diffusion 1.5. D3PO [441] instead treats diffusion generation as the multi-step decision problem. Under mild assumptions, the model is trained by the preference objective at the image segment level. The human annotators are asked about the final image quality and D3PO assumes that any state-action pair of the preferred image is better than that of the rejected image.

4.4 Inference: Enhancing Multimodal Task Performance via Prompt Engineering

After the pretraining and finetuning stages, MMNTP models can also benefit from prompt engineering techniques, much like LLMs. Stemming from research in prompt engineering [387], In-Context Learning (ICL) [94] and Chain-of-Thought reasoning (CoT) [408] are key methods that significantly enhance the performance of LLMs on complex tasks, such as mathematical reasoning [68]. As illustrated in Fig. 16, ICL adds few-shot examples in the prompt of LMM to guide and improve models' performance on unseen examples. CoT guides the model to articulate step-by-step reasoning processes.

Although prompt engineering techniques have had huge success in LLMs [387], their application in multimodal remains largely underexplored so far. Table 4 lists the related work on multimodal ICL and CoT research.

4.4.1 Multimodal In-Context Learning. Multimodal In-Context Learning (ICL) is an emerging paradigm in which models leverage a few demonstration examples incorporating visual, textual, and other optional modalities to perform multimodal tasks. In this learning paradigm, the input processed by the Large Multimodal Model is divided into two components: the query x_q and the

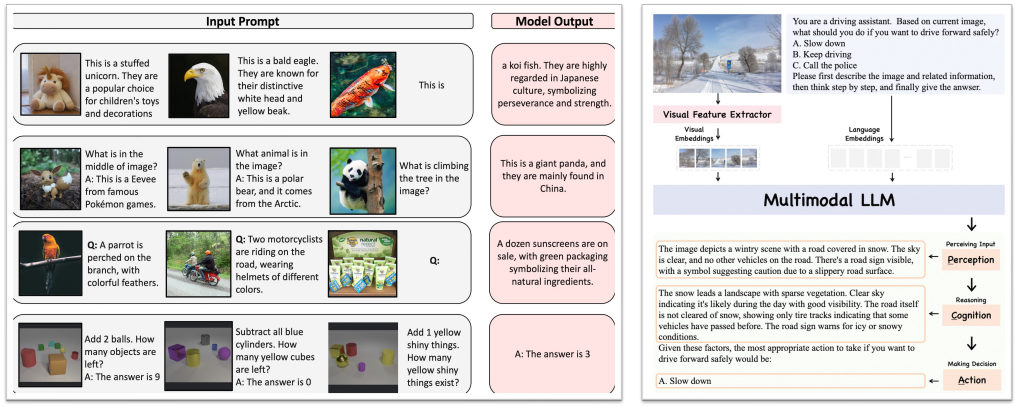

Fig. 15. Examples of Multimodal In-Context Learning and Chain-of-Thought reasoning.

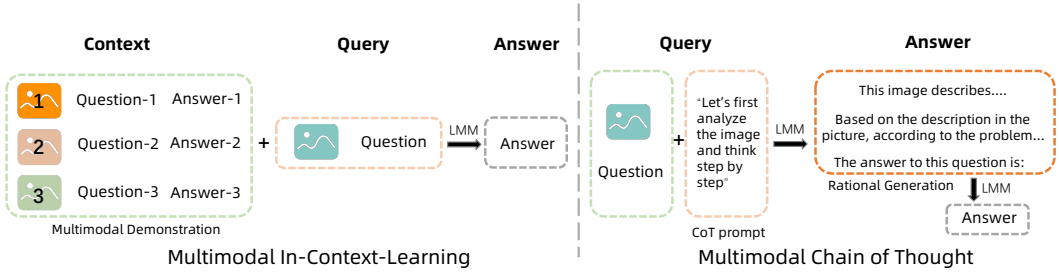

Fig. 16. Multimodal inference enhancement methods.

context C. The LMM needs to generate a sequence of tokens as outputs y_q based on these two parts:

$$y_q = LMM(x_q, C) \tag{12}$$

The context C consists of a set of input-output ICL examples:

$$C = \{(x_i, y_i)\}_{i=1}^{n} \tag{13}$$

Adopting the notation from Todd et al. [375], we represent the generic template for organizing the context C as follows:

$$Q : \{x_1\} \backslash n\, A : \{y_1\} \backslash n \backslash n \ \ldots Q : \{x_n\} \backslash n\, A : \{y_n\}, \tag{14}$$

where Q and A symbolize the question and answer template structures respectively, and x_i and y_i denote the question and answer of the i-th demonstration respectively.

Multimodal ICL introduces unique challenges compared to unimodal ICL, particularly in integrating and aligning diverse modalities such as text, images, and videos [351] [497][15]. In multimodal ICL, both the query x_q and context x_i may vary in modality, conveying complementary yet distinct information that can lead to imbalanced or inefficient learning. A primary challenge, as noted in recent studies [7] [15], is that performance in many multimodal ICL systems remains largely text-driven, with other modalities—such as images or videos—contributing minimally to overall task performance.

Table 4. Summary of Multimodal Prompt Engineering research.

Method	Year	Modality	Backbone Model	Task
Multimodal ICL				
Frozen [382]	2021	Image	GPT2 Architecture [323]	Understanding
Flamingo [3]	2022	Image	GPT2 Architecture [323]	Understanding
MMICL [497]	2023	Image	InstructBLIP [74]	Understanding
EILeV [453]	2023	Image	-	Understanding
Open-Flamingo [7]	2023	Image	Flamingo Architecture [3]	Understanding
LCL [365]	2023	Image	Otter [223], OpenFlamingo [7]	Understanding
Med-Flamingo [296]	2023	Image	Open-Flamingo [7]	Understanding
MIMIC-IT [223]	2023	Image	OpenFlamingo [7]	Understanding
LVM [13]	2023	Image	LLaMA Architecture [378]	Understanding& Generation
LWM [257]	2023	Image, Video	LLaMA Architecture [378]	Understanding & Generation
Yang et al. [444]	2024	Image	Open- Flamingo [7]	Understanding
VisualICL [508]	2024	Image	LLaVA [254]	Understanding
Many-Shots ICL [184]	2024	Image	GPT4-o [302], Gemini1.5 [371]	Understanding
CoBSAT [471]	2024	Image	Emu [362]	Generation
Video ICL [484]	2024	Video	LLaMA Architecture [378]	Generation
Emu [363]	2024	Image, Video	LLaMA [378]	Understanding & Generation
Emu2 [359]	2024	Image, Video	LLaMA-33B [378]	Understanding & Generation
Yang et al. [350]	2024	Image	GPT2 Architecture [323]	Understanding & Generation
VALL-E [391]	2023	Audio	-	Generation
MELLE [289]	2024	Audio	-	Generation
Seed-TTS [4]	2024	Audio	-	Generation
Audio Flamingo [201]	2024	Audio	OPT-IML-MAX-1.3B [173]	Understanding
Moshi [78]	2024	Audio	Helium [78]	Understanding & Generation
Multimodal CoT				
MM-CoT [494]	2023	Image	T5-770M [326]	Understanding
DDCoT [505]	2023	Image	ChatGPT [305]/GPT-3 [34]	Understanding
VCDM [151]	2023	Image	Stable Diffusion [389]	Generation
V* [416]	2024	Image	Vicuna-7B [63]	Understanding
CogCoM [318]	2024	Image	Vicuna-7B [63]	Understanding
VisualCoT [346]	2024	Image	Vicuna-7B/13B [63]	Understanding
CCoT [293]	2024	Image	-	Understanding
VideoCoT [404]	2024	Video	-	Understanding
VoT [116]	2024	Video	Vicuna-7B [63]	Understanding
WavLLM [162]	2024	Audio	LLaMA Architecture [379]	Understanding
SpeechVerse [76]	2024	Audio	Flan-T5-XL [66]	Understanding
CoT-ST [100]	2024	Audio	-	Understanding
AST-CoT [161]	2024	Audio	T5 [326]	Understanding

To address this challenge, several approaches [7, 453, 453, 497] focus on enhancing the model's ability to generalize across diverse multimodal tasks. EILEV [453] proposes new training methods for video understanding. MMICL [497] and CoBSAT [471] use specialized datasets and prompt engineering to enhance multimodal reasoning. Recent work further extends these efforts by exploring large-scale models for more effective in-context learning with interleaved multimodal inputs, [212, 257, 358, 358].

4.4.2 Multimodal Chain-of-Thought Prompting. Multimodal Chain-of-Thought (CoT) is a method that enables models to perform complex reasoning and decision-making in a multimodal setting

through step-by-step derivation and coherent thinking. Pioneered by Zhang et al. [494], MM-CoT introduces Chain-of-Thought prompting into visual domains, raising the challenge of labor-intensive annotation, as multimodal data often demands expensive and complex human-labeled information. MM-CoT employs ScienceQA [273], a dataset focused on scientific questions involving multiple modalities with annotated rationales, while VoT [116] tackles the annotation challenge in video tasks by combining machine and human expertise through active learning.

Another challenge lies in mitigating language hallucinations [3, 53, 179, 285, 330, 488, 499], which are exacerbated due to the lack of necessary and fine-grained visual context when multimodal information is provided simultaneously. To better inject visual information, V* [416] addresses this by dynamically focusing on key visual regions, ensuring that visual details are accurately attended to, particularly in high-resolution images. CCoT [293] generates scene graphs instead of simple captions, explicitly reasoning over visual features to avoid misinterpretation. Moreover, DDCoT [505] introduces a new CoT prompting method that divides the roles of reasoning and visual recognition between language and visual models, thereby enhancing reasoning clarity and reducing hallucinations.

Subsequent work [404] [116] [100] [326] has extended the method beyond images to include video and audio. For instance, the CoT-ST [100]framework adapts chain-of-thought reasoning for speech translation, breaking the process into distinct steps to improve accuracy and fluency. Video-CoT [404] focus on complex video reasoning, aiming to achieve human-level video comprehension.

5 DATASETS AND EVALUATION

In this section, we delve into several crucial aspects of training and evaluating MMNTP models. The subdivision begins with an exploration of the training datasets (Section 5.1), categorized into pre-training and fine-tuning datasets. The pre-training datasets are further divided based on modality into text-only, image-based, video-based, and audio-based data, which are essential for modality alignment and the establishment of a unified multimodal representation. Following this, fine-tuning datasets are described, focusing on their specific applications in multimodal understanding and multimodal generation tasks.

Additionally, we discuss the evaluation of MMMNTP models (Section 5.2), which is pivotal in measuring their effectiveness and capability across various modalities. This aspect is divided into holistic evaluation and emerging evaluation benchmarks. Holistic evaluation benchmarks, such as MME [120] and SEED-Bench [222], comprehensively assess the integration and interplay between different modalities like image, text, and video. Emergent benchmarks, including SparklesEval [167] and HallusionBench [140], push the boundaries further by testing specialized capabilities like conversational competence, mathematical reasoning, and mitigation of hallucinations in model outputs.

5.1 Training Datasets

Depending on the stage of training, we categorize data into pre-training data and fine-tuning data. Pre-training data can be classified into uni-modal data and multimodal data based on modality. Fine-tuning data is categorized based on its usage scenario into multimodal understanding data and multimodal generation data.

5.1.1 Pre-training Datasets. Unlike large language models that are pre-trained only on pure text data, multimodal models require pre-training on a variety of different modalities of data, which demands a significant quantity and diversity of multimodal data. In this section, we briefly summarize several multimodal datasets widely used for training multimodal models. Based on the type of

modality, we categorize these data into four groups: Text-Only, Image-Based, Video-Based, and Audio-Based.

Text-Only. Although pure text data is commonly utilized in language models, it also plays a crucial role in enhancing the language expression and reasoning abilities of multimodal models. For this purpose, pure text data is integrated into the pre-training corpus. One of the most extensively used datasets in this context is C4 [147], a filtered open-source dataset derived from web crawls. Its multilingual variant, mC4 [436], encompasses natural text in 101 languages, sourced from the public Common Crawl web archive. Additionally, the Wikipedia dataset [144], which consists of cleaned articles in multiple languages, is created from language-specific segments of the Wikipedia dump. Another significant contribution to this field is The Pile, an expansive and diverse open-source dataset for language modeling. Amassing a total of 825 GiB, The Pile [125] is an amalgamation of 22 distinct, high-quality smaller datasets, providing a rich resource for language model pretraining. Recently, RedPajama [69], an open dataset with 30 trillion tokens for training large language models, has also been introduced, contributing significantly to the resources available for developing advanced language models. Furthermore, FineWeb [310] release a new, large-scale (15-trillion tokens) dataset for LLM pretraining. FineWeb is derived from 96 CommonCrawl snapshots and produces better-performing LLMs. Dolma [352] is a high-quality open dataset from a diverse mix of web content, academic publications, code, books, and encyclopedic materials, covering 3T tokens.

Image-Based. Multimodal data is key for models to perform modality alignment, that is, to map different modal representations into a unified space. CLIP [321] was developed using 400 million image-text pairs sourced from the internet. Subsequent models like ALIGN [180], BASIC [314], and Florence [464] were trained on even larger and more diverse datasets with noisier image-text pairs. However, the majority of these extensive datasets remain inaccessible to the public. In the academic community, researchers recommend using several million image-text pairs for multimodal model pre-training, including CC12M [45], RedCaps [81], YFCC [372], WIT [355], and Capsfusion [460]. Publicly accessible datasets of a relatively smaller scale include SBU [303], MSCOCO [248], VG [203], and CC3M [347]. Among the larger-scale image-text datasets available to the public are FILIP [446], LAION-400M [345], COYO-700M [35], SA-1B [398], and LAION-5B [344], among others. Additionally, some studies have emphasized the importance of data quality in building robust multimodal models, such as DataComp [123], Shutterstock [299], and ShareGPT4V [50]. Beyond sourcing image-text data from the web, there has been a growing interest in compiling datasets that interleave images and text, a concept pioneered by the M3W [3] dataset featured in Flamingo [3]. Notable examples of such datasets are MMC4 [512] and OBELISC [213]. Additionally, there's an emerging trend in research to focus on the extraction and association of text segments in captions with specific areas in images, leading to the formation of grounded image-text pairs. Datasets like GRIT-20M [313] and CapsFusion-grounded [357] exemplify this methodology.

Video-Based. MSR-VTT [427] features 10K diverse web video clips and 200K clip-sentence pairs spanning a wide range of categories. HowTo100M [292] expands this landscape with 1.22 million YouTube videos on topics like cooking and crafting, enriched with subtitles from ASR systems or manual input. ACAV100M [219] provides a vast 100 million video library, ideal for self-supervised learning with high audio-visual correspondence. WebVid [14] enhances video data with manually crafted, accurate captions. Ego4D [137] offers an extensive collection of diverse egocentric video footage for research. HD-VILA [435] introduces a high-resolution video-language dataset with varied content. YT-Temporal [469], sourced from public YouTube videos, focuses on broadening understanding of objects, actions, and scenes. VideoCC3M [297] utilizes a new pipeline to transfer image captions to videos without extra manual labor. Youku-mPLUG [426] has released the largest

public Chinese video-language dataset, prioritizing safety, diversity, and quality. Most recently, InternVid [403] demonstrates a scalable method for building high-quality video-text datasets using large language models, effectively enhancing video language representation learning.

Audio-Based. Audio-based pretraining datasets can be primarily categorized into three types: speech pretraining datasets, music pretraining datasets, and general audio pretraining datasets. Librilight [190] includes more than 60k hours unlabeled speech data and is widely used by audio pretraining [391, 493]. Libriheavy [191] introduces a refined pipeline for audio alignment and segmentation and detailed annotations with punctuation and capitalization, reflecting more natural speech patterns, to the mostly unlabeled Librilight. Wenetspeech [475] is the largest Mandarin speech pretraining corpus, collecting over 22,400 hours of audio, with 10,000+ hours of high-quality labeled speech, 2,400+ hours of weakly labeled speech, and roughly 10,000 hours of unlabeled speech from diverse sources such as YouTube and podcasts. Yodas [235] offer over 500,000 hours of speech data in more than 100 languages, significantly benefiting the multilingual nature of the audio pretrain community. Other widely-used speech pretraining datasets include librispeech [308], libritts [470] and gigaspeech [48]. Music pretraining is a growing research area [85, 171, 241, 243, 274, 320, 510]. Million Song Dataset (MSD) [23] is one of the largest publicly available collections of audio features and metadata for a million contemporary popular music tracks. FMA (Free Music Archive) Dataset [77] is a well-curated collection of over 100,000 tracks from various artists and genres available under Creative Commons licenses. Other widely-used music pretraining datasets include disco10m [209], mtg-jamendo [25], and Lp-musiccaps [93]. General audio pretraining datasets, including wavcaps [286], audioset [129], vggsound [49], and clotho [98], mainly focus on boosting the performance of localizing audio-visual correspondence and audio-text intermodal translation tasks (not speech-to-text).

5.1.2 *Fine-tuning Datasets.*

Multimodal Understanding. The inaugural work in applying instruction tuning to the multi-modal domain was presented by MultiInstruct [434], which successfully combined multi-modal learning into a single-format benchmark dataset incorporating 62 diverse tasks. Concurrently, LLaVA [254] harnessed the capabilities of the language-centric GPT-4 to generate datasets for multi-modal, instruction-based tasks involving both text and images. MiniGPT-4 [509] precisely assembled a dataset rich in detailed image descriptions to facilitate the convergence of visual and linguistic elements.

Further advancements were marked by LMeye [239], MMEvol [278], PF-1M [46], and SVIT [496], which scaled up the magnitude of instruction tuning. The domains of video content were also explored by Video-Chat [228] and Video-ChatGPT [280], which adapted instruction tuning to this dynamic format. In the specialized medical sector, PMC-VQA [485] and LLaVA-Med [225] crafted datasets for instruction tuning by leveraging existing medical data repositories. Object detection tasks were ingeniously integrated into instruction tuning through the efforts of DetGPT [315] and MGVLID [500]. GPT4Tools [443] was developed to enhance open-source large language models (LLMs) by equipping them with the versatility to utilize an array of tools effectively, while M^3IT expanded the reach of multi-modal instruction tuning across multiple languages. Expanding the horizon further, X-LLM [47], MIMIC-IT [223], MotionGPT [181], Macaw-LLM [279], and BuboGPT [502] ventured into new modalities, enhancing the scope of instruction tuning. The integration of 3D tasks into this domain was initiated by LAMM [449] and M3DBench [232], enriching the complexity and applicability of instruction tuning. Meanwhile, LLaVAR [491] leveraged publicly accessible OCR tools to harvest text-rich images from the LAION [345] dataset, thus enhancing visual instruction tuning processes. To address the phenomenon of hallucinations, HalDetect [142]

Table 5. Statistics of commonly-used Pre-training data.

Datasets	Tags	Doc/Img/Vid/Aud	Source	Time
C4 [147]	Text-Only	8.2M/-/-/-	CommonCrawl	Apr-2019
mC4 [436]	Text-Only	2.1M/-/-/-	CommonCrawl	Oct-2020
Pile [125]	Text-Only	211M/-/-/-	Other	Dec-2020
Wikipedia [144]	Text-Only	13.4M/-/-/-	Wikipedia	Mar-2023
RedPajama [69]	Text-Only	100B/-/-/-	CommonCrawl	Oct-2023
Dolma [352]	Text-Only	4.4B/-/-/-	Common Crawl, GitHub, Reddit, ...	Jan-2024
FineWeb [310]	Text-Only	22.7B/-/-/-	CommonCrawl	May-2024
SBU [303]	Image-Based	1M/1M/-/-	Flickr	Dec-2011
YFCC [372]	Image-Based	100M/99.2M/0.8M/-	Flickr	Jan-2016
MS-COCO [249]	Image-Based	1M/200K/-/-	HumanCurated	Jul-2018
VG [203]	Image-Based	5.4M/108K/-/-	HumanCurated	Feb-2016
CC3M [347]	Image-Based	3.3M/3.3M/-/-	Web Crawl	Jul-2018
CC12M [45]	Image-Based	12M/12M/-/-	Web Crawl	Feb-2021
WIT [355]	Image-Based	37.6M/11.5M/-/-	Wikipedia	Jul-2021
RedCaps [81]	Image-Based	12M/12M/-/-	Reddit links	Nov-2021
FILIP300M [445]	Image-Based	300M/300M/-/-	Web Crawl	Nov-2021
LAION-400M [345]	Image-Based	400M/400M/-/-	CommonCrawl	Nov-2021
Shutterstock [299]	Image-Based	15M/15M/-/-	Shutterstock	Aug-2022
Coyo-700M [35]	Image-Based	747M/747M/-/-	CommonCrawl	Aug-2022
Laion-5B [344]	Image-Based	5B/5B/-/-	CommonCrawl	Oct-2022
DataComp [123]	Image-Based	1.4B/1.4B/-/-	Web Crawl	Apr-2023
SA-1B [398]	Image-Based	1.1B/11M/-/-	Photo Company	Aug-2023
Capsfusion [460]	Image-Based	120M/120M/-/-	Other	Oct-2023
ShareGPT4V [50]	Image-Based	1.2M/1.2M/-/-	Other	Nov-2023
M3W [3]	Image-Based (Interleaved)	-/185M/-/-	Web Crawl	Apr-2022
MMC4 [512]	Image-Based (Interleaved)	103M/585M/-/-	Other	Apr-2023
Obelisc [213]	Image-Based (Interleaved)	141M/353M/-/-	Web Crawl	Jun-2023
GRIT-20M [313]	Image-Based (Grounded)	20M/20/-/-	Other	Jun-2023
CapsFusion-grounded [357]	Image-Based (Grounded)	100M/100M/-/-	Other	Dec-2023
MSR-VTT [428]	Video-Based	200K/-/10K/-	HumanCurated	Jun-2016
HowTo100M [292]	Video-Based	136M/-/1.2M/-	Youtube	Jun-2019
ACAV [219]	Video-Based	-/-/100M/100M	Web Crawl	Jan-2021
WebVid [14]	Video-Based	10M/-/10M/-	Stock Footage	Jan-2021
Ego4D [137]	Video-Based	-/-/-/-	HumanCurated	Oct-2021
HD-VILA [435]	Video-Based	100M/-/3.3M/-	YouTube	Nov-2021
YT-Temporal [469]	Video-Based	1B/-/20M/-	YouTube	Jan-2022
VideoCC3M [297]	Video-Based	10.3M/-/6.3M/-	Other	Apr-2022
Youku-mPLUG [426]	Video-Based	10M/-/10M/-	Youku	Jun-2023
InternVid [403]	Video-Based	234M/-/7.1M/-	YouTube	Jul-2023
Million Song Dataset [23]	Audio-Based	-/-/-/1M	The Echo Nest	Feb 2011
MTT [215]	Audio-Based	-/-/-/25.8k	Web Crawl	June 2013
LibriSpeech [308]	Audio-Based	155.8k/-/-/1k hours	Audio Books	Jun 2015
FMA [77]	Audio-Based	-/-/-/106k	Free Music Archive	Dec 2016
Audio Set [129]	Audio-Based	2.1M/-/-/2.1M	YouTube	Mar-2017
LibriTTS [470]	Audio-Based	2.4k/-/-/0.58k hours	Audio Books	Apr 2019
MTG-Jamendo [25]	Audio-Based	-/-/-/55k	Jamendo	Jun 2019
Clotho [98]	Audio-Based	25k/-/-/5k	FreeSound Platform	Oct 2019
Librilight [190]	Audio-Based	-/-/-/60k hours	Audio Books	Dec 2019
VGGSound [49]	Video-Based	309/-/200k/200k	Web Crawl	Apr 2020
Gigaspeech [48]	Audio-Based	-/-/-/40k hours	Audio Books	Jun 2021
LAION-Audio-630k [418]	Audio-Based	630k/-/-/630k	Web Crawl	Nov 2021
wenetspeech [475]	Audio-Based	-/-/-/22.4k hours	Youtube	Feb 2022
WavCaps [286]	Audio-Based	400k/-/-/400k	Other	Mar-2023
LP-MusicCaps [93]	Audio-Based	2.2M/-/-/0.5M	Web Crawl	Jul 2023
LibriHeavy [191]	Audio-Based	9M/-/-/50k hours	Audio Books	Sep 2023
disco-10m [209]	Audio-Based	-/-/-/15.2M	Youtube	2023
yodas [235]	Audio-Based	-/-/-/500k hours	Youtube	Dec 2023

developed a pioneering multi-modal dataset focused on accurate image descriptions. In the pursuit of robustness, GAVIE [250] introduced a mix of positive and negative instructions, fortifying the training for visual instruction tuning. StableLLaVA [244] combined the generative prowess of ChatGPT with text-to-image models to produce a versatile and diversified dataset featuring a wide range of image content. Sparkles [167] introduced the first machine-generated dialogue dataset tailored for word-level interleaved multi-image and text interactions. The project LVIS-INSTRUCT4V [393]

Table 6. Statistics of commonly-used instruction tuning data.

Datasets	Tags	Nums	Source	Time
MultiInstruct [434]	Image+Text	235K	Existing datasets + Human	Dec-2022
LLaVA [254]	Image+Text	158K	COCO + GPT	April-2023
Mini-GPT4 [509]	Image+Text	3.5K	CC3M + GPT	April-2023
LMeye [239]	Image+Text	7.3M	Existing datasets + GPT	May-2023
X-LLM [47]	Image+Video+Audio+Text	10K	Existing datasets + GPT	May-2023
Video-Chat [228]	Video+Audio+Text	11K	WebVid-10M + GPT	May-2023
PMC-VQA [485]	Image+Text	227K	PMC-OA + GPT	May-2023
DetGPT [315]	Image+Text	30K	COCO+GPT	May-2023
GPT4Tools [443]	Image+Text	71K	Visual ChatGPT	May-2023
LLaVA-Med [225]	Image+Text	60K	PubMed +GPT	June-2023
M^3IT [231]	Image+Text	2.4M	Existing datasets + GPT	June-2023
MIMIC-IT [223]	Image+Video+Text	2.8M	Existing datasets + GPT	June-2023
Video-ChatGPT [280]	Video+Text	100K	ActivityNet-200+Human	June-2023
LAMM [449]	Image+Text	196K	Existing datasets + GPT	June-2023
LLaVAR [491]	Image+Text	422K	LAION-5B + GPT	June-2023
Macaw-LLM [279]	Image+Video+Audio+Text	119K	Existing datasets + GPT	June-2023
GAVIE [250]	Image+Text	400K	Existing datasets + GPT	June-2023
MotionGPT [181]	Motion+Text	50K	Existing datasets+Human	July-2023
PF-1M [46]	Image+Text	1M	Existing datasets+GPT	July-2023
SVIT [496]	Image+Text	4.2M	Existing datasets + GPT	July-2023
BuboGPT [502]	Image+Audio+Text	170K	Existing datasets + GPT	July-2023
MGVLID [500]	Image+Text	108K	Existing datasets + GPT	July-2023
HalDetect [142]	Image+Text	16K	COCO+Human	Aug-2023
StableLLaVA [244]	Image+Text	126K	SD+GPT	Aug-2023
Sparkles [167]	Image+Text	6.5K	Existing datasets + GPT	Aug-2023
LVIS-INSTRUCT4V [393]	Image+Text	220K	Existing dataset+GPT	Nov-2023
M3DBench [232]	Image+Text	320K	Existing datasets + GPT	Dec-2023
MMEvol [278]	Image+Text	480K	Existing datasets + GPT	Sept-2024
InstructPix2Pix [33]	Image Editing	313K	SD+GPT	Jan-2023
HIVE [483]	Image Editing	1.1M	Existing datasets + SD + GPT	Mar-2023
MagicBrush [481]	Image Editing	10K	Existing datasets + SD + Human	Nov-2023
HQ-Edit [170]	Image Editing	200K	SD+GPT	Apr-2024
UltraEdit [498]	Image Editing	4.1M	Existing datasets +SD+GPT	June-2024

capitalized on the improved visual processing strengths of GPT-4 to achieve a higher precision in image detail capture and instruction annotation accuracy.

Multimodal Generation. Additionally, some instruction-based image editing datasets focus on image generation. A typical dataset is InstructPix2Pix [33], which initially uses GPT-3 [34] to generate the text for edits, and then utilizes Stable Diffusion [336] along with Prompt2Prompt [156] technology to generate the corresponding edited images to construct the dataset. Furthermore, HIVE [483] introduces a larger number of training triplets and incorporates human ranking results, providing stronger supervision signals for more effective model training. Building on these advancements, MagicBrush [481] introduces the first large-scale, manually annotated dataset specifically designed for instruction-guided real image editing. Expanding further, HQ-Edit [170] provides a high-quality instruction-based image editing dataset consisting of approximately 200,000 edits. Unlike previous methods that relied on attribute guidance or human feedback to build datasets, HQ-Edit employs a scalable data collection pipeline that leverages advanced foundation models, specifically GPT-4V and DALL-E 3.

5.2 Evaluation

The evaluation MMNTP models is crucial to understand their capabilities, limitations, and potentials across different dimensions. This section delves into the different facets of evaluating such models, outlining both established holistic benchmarks and emerging evaluation practices.

Table 7. Statistics of Benchmark.

Benchmark	Modalities	Samples	Span / Feature	Release Date
Holistic Evaluation				
MME [120]	text, image	2,374	14 Subtasks	2023-06-23
MMBench [261]	text, image	4,377	20 Dimensions	2023-07-12
SEED-Bench [222]	text, image	19,242	12 Dimensions	2023-07-30
MLLM-Bench [126]	text, image	420	6 Dimensions	2023-11-23
MMMU [466]	text, image	11,550	183 Subfields	2023-11-27
MVBench [229]	text, video	4,000	20 Subtasks	2023-11-28
SEED-Bench-2 [221]	text, image, video	24,000	27 Dimensions	2023-11-28
VBench [168]	text, video	1,600	16 Dimensions	2023-11-29
CMMMU [478]	text, image	12,000	30 Subjects	2024-01-22
Emerging Benchmarks				
SparklesEval [167]	text, image	1,967	Multi-modal Dialogue	2023-08-31
MathVista [272]	text, image	6,141	Math Reasoning	2023-10-03
HallusionBench [140]	text, image	1,129	Hallucination	2023-10-23
Bingo [73]	text, image	370	Hallucination	2023-11-06
MMC-Benchmark [251]	text, image	2,126	Chart Reasoning	2023-11-15
BenchLMM [37]	text, image	1,967	Style Robustness	2023-12-05
TVGE [414]	text, video	2,543	New Metric	2024-01-15
MMCBench [480]	text, image, speech	4,000	Self-consistency	2024-01-22
VQAv2-IDK [42]	text, image	6,624	Hallucination	2024-02-15
PCA-Bench [52, 53]	text, image	1200	Embodied-AI	2024-02-21
MATH-Vision [394]	text, image	3,040	Math Reasoning	2024-02-22
TempCompass [262]	text, video	7,540	Video Understanding	2024-03-01
MMEvalPro [163]	text, image	2,138	Reasoning, Calibration	2024-06-29

5.2.1 Holistic Evaluation. In the assessment of multi-modal large language models, holistic benchmarks serve as foundational tools for evaluating the integration and interplay between different modalities such as image, text, and video.

Within the domain of image-language, benchmarks like MME [120] offer a comprehensive evaluation of models' perception and cognition abilities across a diverse set of tasks, emphasizing the importance of intuitive and quantifiable analysis without the need for extensive prompt engineering. MMBench [261] extends this by incorporating a vast dataset and a unique evaluation strategy, CircularEval, to robustly test models across a wide array of capabilities, including object localization and social reasoning, through single-choice questions derived from a broad spectrum of ability dimensions. SEED-Bench [222] and its successor SEED-Bench-2 [221] further contribute by providing a detailed assessment framework that covers generative comprehension capabilities across various dimensions, utilizing a mix of automatic filtering and manual verification to ensure the relevance and quality of questions. MLLM-Bench [126] aims to reflect user experiences more accurately by focusing on diverse scenarios ranging from perception to creative output, highlighting the gaps in performance between existing models and suggesting directions for future development. MMMU [466] uniquely challenges models on college-level subject knowledge across a wide range of disciplines, requiring advanced perception and reasoning over complex multi-modal questions. CMMMU [478] is designed to assess the proficiency of multimodal models in Chinese, featuring 12,000 questions across six disciplines and 30 subjects. It challenges models with complex reasoning tasks and a variety of image types.

In the video-language category, benchmarks like MVBench [229] specifically target the temporal understanding capabilities of models by focusing on dynamic, video-based reasoning tasks that

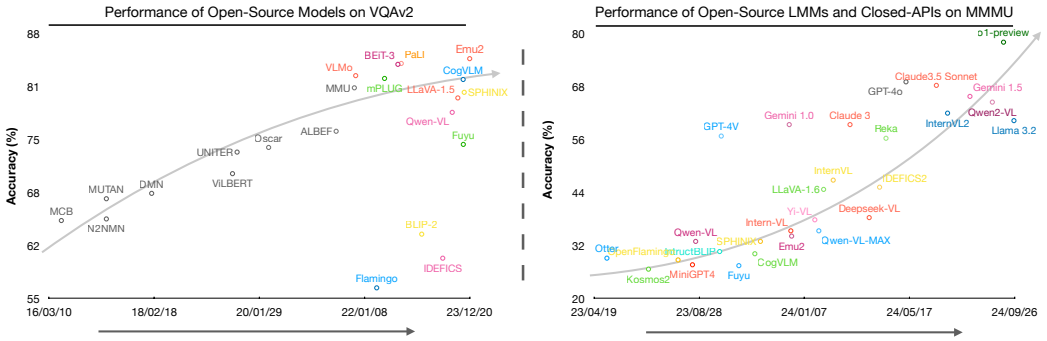

Fig. 17. Performance evaluation of various models on the VQAv2 [136] and MMMU [466] benchmarks. Colored representations signify the employment of the next token prediction architecture, whereas gray depictions denote alternative architectural frameworks.

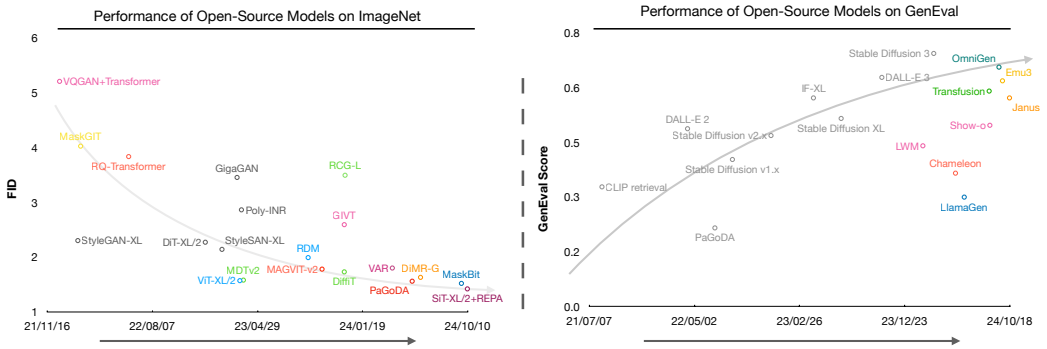

Fig. 18. Performance evaluation of various models on the ImageNet [342] and GenEval [131] benchmarks. Colored representations signify the employment of the next token prediction architecture, whereas gray depictions denote alternative architectural frameworks.

extend beyond static image understanding. This involves evaluating models on their ability to interpret action sequences, object interactions, and scene transitions within video content. VBench [168] offers a nuanced approach to assessing video generation quality by breaking down the evaluation into specific dimensions and providing detailed feedback on models' performance across various content types, thereby enhancing our understanding of video generative models.

5.2.2 Emerging Evaluation Benchmarks. Emerging benchmarks delve into more specialized and advanced aspects of multi-modal understanding, pushing the boundaries of model assessment. SparklesEval [167] focuses on conversational competence in multi-modal contexts, emphasizing the ability of models to maintain coherent conversations involving multiple images and dialogue turns. MathVista [272] challenges models on their mathematical reasoning abilities within visual contexts, incorporating a wide range of tasks that require a blend of visual understanding and compositional reasoning. HallusionBench [140] is designed to test models on their ability to handle nuanced visual interpretations, particularly in the context of image-context reasoning, while Bingo [73] addresses the critical issue of hallucinations in models, focusing on understanding and quantifying biases and interference effects. MMC-Benchmark [251] stands out for its focus on chart understanding,

offering a unique set of tasks that evaluate models' abilities to extract and reason with information from visual charts, marking a significant challenge for even advanced models. BenchLMM [37] assesses performance across different visual styles, crucial for understanding and improving visual reasoning capabilities in diverse real-world scenarios. Lastly, TVGE [414] introduces a novel metric, the Text-to-Video Score (T2VScore), for evaluating text-to-video generation models, providing a comprehensive tool for assessing alignment with textual descriptions and overall video quality. MMCBench [480] is designed to evaluate LMMs robustness and self-consistency across text, image, and speech modalities, focusing on four generation tasks: text-to-image, image-to-text, text-to-speech, and speech-to-text. The purpose of VQAv2-IDK [42] is to challenge and evaluate models on their ability to recognize and admit uncertainty or lack of information in visual question answering tasks, rather than generating incorrect or arbitrary responses. Math-Vision [394] benchmark is a comprehensive dataset of 3,040 mathematical problems with visual contexts, spanning 16 subjects and 5 difficulty levels, aimed at evaluating the reasoning capabilities of LLMs in mathematical scenarios. TempCompass [262] benchmark assesses Video LLMs' understanding of temporal dynamics in videos through diverse tasks and formats, highlighting significant gaps in models' ability to perceive time, using an LLM-based automatic evaluation method. MathVerse [482] benchmark offers varying degrees of textual and image information content in multi-modality math problems, contributing to 2,612 test samples in total to investigate the ability of VLMs to gain information from pictures.

These holistic and emerging benchmarks provide a comprehensive framework for evaluating the current capabilities and identifying the limitations of multi-modal large language models, guiding the path towards more sophisticated, versatile, and capable multi-modal AI systems.

6 CHALLENGES

In this section, we propose four currently unsolved challenges, primarily stemming from the MMNTP training paradigm. We also recommend that the readers refer to surveys that address other open challenges, such as evaluation of multimodal LLMs [121, 164], efficient LMM architectures [185], generative approaches and auto-regressive models for vision [182].

6.1 Scaling Up MMNTP Models with Unlabeled Multimodal Data

The utilization of abundant unlabeled text data is one key to the success of LLMs [501]. However, the potential of using large amounts of unlabeled multimodal data in training MMNTP models has not been fully investigated. Most large multimodal models currently rely on labeled pair-wise data, such as image-caption or audio-text pairs during training. Recent studies [277, 369, 438] have attempted to leverage the enormous amounts of interleaved text-image data available on the internet. However, their performance on downstream tasks does not provide a significant advantage over models trained solely on pair-wise data, such as LLaVA [254] and Qwen2-VL [395]. Consequently, determining how to better utilize unlabeled multimodal data such as text-image interleaved webpages, rich-text figures, text-free audios, videos, screenshots of graphical user interface (GUI), etc., remains a crucial question in the effort to scale up MMNTP models.

From another perspective, the benefits from scaling up unlabeled text data and model sizes can be curved with Scaling Law [159, 192], which forms a fundamental basis and faith for the development of LLMs. It elucidates the intricate relationship between model performance, model size, and the amount of training data, while also guiding the optimal allocation of computational resources during LLM training. However, the return of scaling MMNTP models remains largely under-explored. Some studies [2, 361, 373] have explored or hypothesized the scaling behaviors of MMNTP models in the self-supervised training manner. According to Aghajanyan et al. [2], data from distinct modalities exhibit varying scaling behaviors. However, the reasons behind these

differences and their impact on the performance of downstream tasks across various modalities remain unclear. Furthermore, it is still uncertain whether MMNTP models can develop similar emergent abilities [407] in downstream tasks as LLMs do when the training is scaled up.

6.2 Overcome Modalities' Interference and Boost Synergy in MMNTP Training

The second challenge comes with the multitask learning nature of MMNTP models, where predicting tokens belonging to different modalities could be viewed as different tasks. A critical challenge within this framework is maintaining the performance of each individual task from interference while investigating whether tasks from different modalities can provide mutual assistance [71].

A primary concern here is modality interference [490], where the performance of one task might negatively impact another due to conflicting information or noise from different modalities. Recent studies underscore how jointly training multimodal tasks in an NTP fashion can pose optimization challenges, especially when a single transformer decoder model is used to generate both text and image outputs [369]. To mitigate issues like gradient norm explosion, techniques such as QK-Norm have been employed [155]. However, the root causes of these optimization difficulties in MMNTP models remain largely unexplored. Further evidence of interference is seen when MMNTP models are built on pretrained large language models, as the language capability often deteriorates when adding more modalities [490]. This demonstrates that interference can lead to suboptimal learning outcomes, as the model struggles to balance competing demands from the different modalities.

6.3 Increase Efficiency in the Training and Inference of MMNTP Models

Efficiency remains a long-standing goal in the training and deployment of deep learning models [290]. Due to the similarities in backbone model structures and NTP training objectives between LLMs and MMNTP models, many advanced methods designed to enhance the efficiency of LLMs [390] can also be effectively applied to MMNTP models. However, there are also several new challenges arising in the training and inference of MMNTP models due to the involvement of data from different modalities.

Training System Efficiency. A big challenge in scaling up MMNTP models lie in the low efficiency in training large-scale MMNTP models on massive GPUs due to the inherent heterogeneity of both models and data across different modalities. Different from the text (1D) data in LLMs, MMNTP training involves high dimensional representations such as image (2D) and video (3D) data, where little has been done to optimize the training of these models from a system perspective. In particular, MMNTP exhibits scaling dependence with modality encoders. Recent studies have found that substantial idle GPU time (GPU Bubble) arises from the complex data dependencies when training MMNTP models that employ both visual encoders and a backbone LLM with pipeline parallelism [298]. To address the issue, several efficient and adaptive training frameworks have been developed to optimize the scheduling of encoder computations and overlap GPU communication with computation. For example, Optimus reduces training time by decomposing image encoder layer computations into smaller kernels and scheduling those kernel executions within LLM bubbles, minimizing the pipeline idle time [117]. DistTrain leverages disaggregated model orchestration and data reordering to improve the training efficiency and scalability of MMNTP, achieving significant improvements in model FLOPS utilization and throughput [495]. Despite these advancements, the unique challenges posed by multimodal architecture, how to develop more advanced system optimizations to train MMNTP on large-scale production clusters with thousands of GPUs remains an open research question.

MMNTP training also exhibit highly variable sequence lengths. As MMNTP models move towards more complex tasks such as multi-image reasoning, multi-modal RAG, video understanding,

supporting MMNTP with long and variable sequence length becomes critical. However, existing large model training systems and the underlying parallelism technologies (data, tensor, pipeline) are limited in their ability to support efficient long sequence training. Recent studies have proposed several sequence parallelism techniques, such as DeepSpeed-Ulysses [174], Ring-Attention [259], and Unified Sequence Parallelism [114], to enable long sequence training for LLMs. However, applying these sequence parallelism strategies to MMNTP needs to take careful consideration in handling heterogeneous data from different modalities, each with distinct characteristics and sequence lengths, which motivates advanced system-algorithm co-design to address this challenge.

Multimodal Tokenization Efficiency in MMNTP Models. As mentioned in the tokenization section, multimodal input like image, audio and video originally resides in a continuous space, which contains a lot of redundant information. The multimodal tokenization process has large room for efficiency improvement, where the core question is: can we use less tokens to represent the multimodal input while maintaining the performance? In the scope of single-modal image modeling and dual-encoder VLM architectures, various model compression methods—such as pruning, knowledge distillation, and quantization—are applied to accelerate image encoders. Model pruning [246, 511] sparsifies the encoder backbone and removes certain modules. Knowledge distillation [415, 465] utilizes soft labels as supervision and train competent smaller dense models. Model quantization [267, 462] replaces models and computation to low-precision counterparts. Though proven effective in single-model scenarios, the adaptability of these methods to training MMNTP models remains largely unexplored.

Modeling Efficiency in Understanding and Generation. Although image tokens occupy a significant portion of the input sequence of MMNTP models, it is discovered that the the LLM backbone only pays a small portion of attention to the image tokens compared to the language tokens [54]. This phenomenon raises the question whether we can reduce the number of image tokens during training and inference without sacrificing performance as they take up most of the sequence length but gain the least attention from the model. In single-modal image modeling, these tokens can be largely pruned using pre-trained priors [284, 425] or through entirely training-free methods [26], with minimal impact on performance. For MMNTP models such as Llava [254] and QwenVL [12], Chen et al. [54] proposed a pruning approach that removes most image tokens without compromising performance. However, the reason behind the redundancy of image tokens and how to leverage this phenomenon remains underexplored for MMNTP models of different modalities.

In the realm of multimodal generation, the challenge of modeling efficiency is equally pronounced. A central issue is how to define an effective generative training objective that suits the Next Token Prediction (NTP) manner for various modalities, given that data from different modalities possess distinct structures. A vanilla NTP training objective may not adequately address these differences, prompting the development of specialized objectives tailored to the characteristics of each modality. For instance, methods like MaskGIT [44], MAGViT [454], VAR [373], and DnD-Transformer [51] have been proposed to better accommodate the unique aspects of different modalities. Moreover, enhancing generation quality efficiently can be achieved through post-generation refinement techniques such as super-resolution [270]. Super-resolution methods aim to upscale the outputs of language models by either fine-tuning the backbone model [91, 92] or by incorporating additional modules [199, 452]. Despite significant advancements in this area, diffusion models remain the de facto model in visual generation applications. However, a detailed comparison of generation quality and efficiency between MMNTP models and diffusion models is still lacking in the literature, leaving room for further exploration and research.

6.4 MMNTP as Universal Interfaces

LLMs have exhibited notable advancements in collaborating with external models in the framework of NTP [145, 150, 319, 349]. It highlights the growing potential for language models to extend their capabilities beyond mere text generation. In this survey, we have mentioned that the next token prediction paradigm has been unifying vision, audio and different multimodal task. However, an ultimate challenge is beyond current explored modalities, achieving a universal interface connecting tasks from various sources, such as robotics [32], molecular [119] and proteins [341]. The key problem is how to formulate a different task as next token prediction, and whether such formulation is efficient and scalable in solving the problem, which is largely underexplored outside the language, vision and audio data.

Design NTP Training Objectives for Different Modalities. For non-text modalities, simply linearizing the data into a 1D sequence and conducting NTP (Next Token Prediction) training may not be the most effective approach. This strategy poses two potential issues: it can overlook the inherent structure of multimodal data and lead to excessively long sequence lengths. For instance, the spatial relationships in images and temporal relationships in videos are crucial elements that the traditional NTP training fails to consider. The number of tokens required to represent an image or video increases linearly with the image's resolution and the video's duration. To address these challenges, several approaches have been developed to adapt NTP training objectives to various modalities, such as images and videos. Notable examples include MaskGiT [44], VAR [373], DnD-Transformer [51], and Next-Block-Prediction [5]. These efforts aim to better capture the unique structures present in different types of data and can reduce the inference time by generating multiple tokens at one time.

Comparison to Diffusion. Diffusion models represent another popular framework for generative modeling, and they have been extensively applied to multimodal data beyond the original image domain [442]. These applications extend to areas such as language [236], robotics [62], and drug discovery [169]. When comparing NTP and diffusion models, both approaches share the fundamental idea of breaking down a complex generation task into multiple, more manageable steps. The major difference between these two approaches lies in how the task is decomposed. NTP breaks down the data according to its dimensional order, whereas diffusion models deconstruct the data globally in a coarse-to-fine manner. There is no consensus on which method is superior, and new approaches are emerging that combine these two modeling techniques. For instance, auto-regressive diffusion models like MAR and Transfusion [234, 507] incorporate elements of both strategies. Exploring how different modeling methods perform on various modalities is a fascinating frontier in the field.

7 CONTRIBUTIONS AND ACKNOWLEDGMENTS

Liang Chen leads the project. **Zekun Wang**, **Shuhuai Ren**, **Lei Li**, **Haozhe Zhao**, and **Yunshui Li** are the core contributors who help drafted the initial version of this survey. **Hongcheng Guo** contributes to the general structure and tokenizer part. **Lei Zhang** contributes to the Datasets and Evaluation section. **Lingwei Meng**, **Shujie Hu**, and **Ge Zhang** contribute to the audio part. **Zefan Cai**, **Yichi Zhang** and **Ruoyu Wu** contribute to the Inference Enhancement section. **Yizhe Xiong** and **Minjia Zhang** contributes to the challenge section. **Qingxiu Dong** contributes to the In-Context Learning part and provides valuable feedback on survey writing. **Yulong Chen**, **Andreas Vlachos**, **Xu Tan**, **Junyang Lin**, **Jian Yang**, **Shuai Bai**, **Wen Xiao**, **Aaron Yee** and **Tianyu Liu** contribute to the revision and discussion throughout the writing of the survey. **Baobao Chang** is the supervisor of **Liang Chen**. We sincerely thank **Sanyuan Chen** (Meta), **Yuchen Yang** (PKU) for their insightful and valuable feedback. The survey will be updated periodically, we also kindly welcome all kinds of comments and suggestions.

REFERENCES

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[2] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. 2023. Scaling Laws for Generative Mixed-Modal Language Models. arXiv:2301.03728 [cs.CL] https://arxiv.org/abs/2301.03728

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[4] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. *arXiv preprint arXiv:2406.02430* (2024).

[5] Anonymous. 2024. Next Block Prediction: Video Generation via Semi-Auto-Regressive Modeling. In *Submitted to The Thirteenth International Conference on Learning Representations.* https://openreview.net/forum?id=JUYBEmwSJK under review.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. arXiv:2103.15691 [cs.CV]

[7] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *ArXiv preprint* abs/2308.01390 (2023).

[8] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Foundational Models Defining a New Era in Vision: A Survey and Outlook. arXiv:2307.13721 [cs.CV] https://arxiv.org/abs/2307.13721

[9] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *arXiv preprint arXiv: 2202.03555* (2022).

[10] Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453* (2019).

[11] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[12] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *ArXiv preprint* abs/2308.12966 (2023).

[13] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. 2023. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785* (2023).

[14] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision.*

[15] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What Makes Multimodal In-Context Learning Work?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition. 1539–1550.

[16] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (Feb. 2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[17] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv: 2106.08254* (2021).

[18] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our Multimodal Models. https://www.adept.ai/blog/fuyu-8b

[19] Gulcin Baykal, Melih Kandemir, and Gozde Unal. 2024. EdVAE: Mitigating Codebook Collapse with Evidential Discrete Variational Autoencoders. arXiv:2310.05718 [cs.CV] https://arxiv.org/abs/2310.05718

[20] Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S. Song. 2024. Genomic Language Models: Opportunities and Challenges. arXiv:2407.11435 [q-bio.GN] https://arxiv.org/abs/2407.11435

[21] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. arXiv:1308.3432 [cs.LG] https://arxiv.org/abs/1308.3432

[22] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In *ICML*, Vol. 2. 4.

[23] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. (2011).

[24] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. 2022. FlexiViT: One Model for All Patch Sizes. *arXiv preprint arXiv: 2212.08013* (2022).

[25] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. ICML.

[26] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token Merging: Your ViT But Faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=JroZRaRw7Eu

[27] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. 2024. An Introduction to Vision-Language Modeling. arXiv:2405.17247 [cs.LG] https://arxiv.org/abs/2405.17247

[28] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing* 31 (2023), 2523–2533.

[29] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39 (1952), 324. https://api.semanticscholar.org/CorpusID:125209808

[30] Andrew Brock, Soham De, Samuel L. Smith, and K. Simonyan. 2021. High-Performance Large-Scale Image Recognition Without Normalization. *International Conference on Machine Learning* (2021).

[31] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. 2021. High-Performance Large-Scale Image Recognition Without Normalization. arXiv:2102.06171 [cs.CV] https://arxiv.org/abs/2102.06171

[32] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818 [cs.RO]

[33] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

[34] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information*

Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[35] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

[36] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. arXiv:2402.12451 [cs.CV] https://arxiv.org/abs/2402.12451

[37] Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. 2023. BenchLMM: Benchmarking Cross-style Visual Capability of Large Multimodal Models. arXiv:2312.02896 [cs.CV]

[38] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2021. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv:2105.05537 [eess.IV]

[39] Shiyue Cao, Yueqin Yin, Lianghua Huang, Yu Liu, Xin Zhao, Deli Zhao, and Kaiqi Huang. 2023. Efficient-VQGAN: Towards High-Resolution Image Generation with Efficient Vision Transformers. arXiv:2310.05400 [cs.CV]

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294

[41] Joao Carreira, Skanda Koppula, Daniel Zoran, Adria Recasens, Catalin Ionescu, Olivier Henaff, Evan Shelhamer, Relja Arandjelovic, Matt Botvinick, Oriol Vinyals, Karen Simonyan, Andrew Zisserman, and Andrew Jaegle. 2022. HiP: Hierarchical Perceiver. *arXiv preprint arXiv: 2202.10890* (2022).

[42] Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. 2024. Visually Dehallucinative Instruction Generation: Know What You Don't Know. arXiv:2402.09717 [cs.CV]

[43] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. arXiv:2301.00704 [cs.CV]

[44] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, T Freeman, and Google Research. [n. d.]. MaskGIT: Masked Generative Image Transformer. ([n. d.]).

[45] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3558–3568.

[46] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003* (2023).

[47] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160* (2023).

[48] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909* (2021).

[49] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

[50] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793* (2023).

[51] Liang Chen, Sinan Tan, Zefan Cai, Weichu Xie, Haozhe Zhao, Yichi Zhang, Junyang Lin, Jinze Bai, Tianyu Liu, and Baobao Chang. 2024. A Spark of Vision-Language Intelligence: 2-Dimensional Autoregressive Transformer for Efficient Finegrained Image Generation. arXiv:2410.01912 [cs.CV] https://arxiv.org/abs/2410.01912

[52] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2023. Towards End-to-End Embodied Decision Making via Multi-modal Large Language Model: Explorations with GPT4-Vision and Beyond. *ArXiv* (2023).

[53] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. 2024. PCA-Bench: Evaluating Multimodal Large Language Models in Perception-Cognition-Action Chain. arXiv:2402.15527 [cs.CL]

[54] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. arXiv:2403.06764 [cs.CV]

[55] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. 2020. Generative Pretraining from Pixels. (2020).

[56] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[57] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058* (2022).

[58] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=mWVoBz4W0u

[59] Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. 2024. A Single Transformer for Scalable Vision-Language Modeling. arXiv:2407.06438 [cs.CV] https://arxiv.org/abs/2407.06438

[60] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238* (2023).

[61] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. arXiv:2312.14238 [cs.CV] https://arxiv.org/abs/2312.14238

[62] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. arXiv:2303.04137 [cs.RO] https://arxiv.org/abs/2303.04137

[63] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

[64] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*. PMLR, 3915–3924.

[65] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919* (2023).

[66] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53. http://jmlr.org/papers/v25/23-0870.html

[67] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 244–250.

[68] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] https://arxiv.org/abs/2110.14168

[69] Together Computer. 2023. *RedPajama: an Open Dataset for Training Large Language Models*. https://github.com/togethercomputer/RedPajama-Data

[70] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems* 36 (2024).

[71] Michael Crawshaw. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. arXiv:2009.09796 [cs.LG] https://arxiv.org/abs/2009.09796

[72] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2023. A Survey on Multimodal Large Language Models for Autonomous Driving. arXiv:2311.12320 [cs.AI] https://arxiv.org/abs/2311.12320

[73] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges. arXiv:2311.03287 [cs.LG]

[74] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv preprint* abs/2305.06500 (2023).

[75] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv:2205.14135 [cs.LG] https://arxiv.org/abs/2205.14135

[76] Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. 2024. SpeechVerse: A Large-scale Generalizable Audio Language

Model. *arXiv preprint arXiv:2405.08295* (2024).
[77] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2016. FMA: A dataset for music analysis. *arXiv preprint arXiv:1612.01840* (2016).
[78] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037* (2024).
[79] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, J. Heek, Matthias Minderer, Mathilde Caron, A. Steiner, J. Puigcerver, Robert Geirhos, Ibrahim M. Alabdulmohsin, Avital Oliver, Piotr Padlewski, A. Gritsenko, Mario Luvci'c, and N. Houlsby. 2023. Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution. *Neural Information Processing Systems* (2023). https://doi.org/10.48550/arXiv.2307.06304
[80] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. arXiv:2409.17146 [cs.CV] https://arxiv.org/abs/2409.17146
[81] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431* (2021).
[82] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems* 36 (2023), 18090–18108.
[83] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143* (2020).
[84] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
[85] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).
[86] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG] https://arxiv.org/abs/2105.05233
[87] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. 2024. Unveiling Encoder-Free Vision-Language Models. arXiv:2406.11832 [cs.CV] https://arxiv.org/abs/2406.11832
[88] Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, and Jiawei Wang. 2023. Write and Paint: Generative Vision-Language Models are Unified Modal Learners. In *The Eleventh International Conference on Learning Representations*.
[89] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. 2017. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. *arXiv preprint arXiv: 1711.08200* (2017).
[90] Sander Dieleman. 2022. Diffusion models are autoencoders. https://benanne.github.io/2022/01/31/diffusion.html
[91] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Dong Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering Text-to-Image Generation via Transformers. *Neural Information Processing Systems,Neural Information Processing Systems* (Dec 2021).
[92] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. [n. d.]. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. ([n. d.]).
[93] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372* (2023).
[94] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A Survey for In-context Learning. arXiv:2301.00234 [cs.CL]
[95] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2023. DreamLLM: Synergistic Multimodal Comprehension and Creation. *arXiv preprint arXiv: 2309.11499* (2023).
[96] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[97] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An Embodied Multimodal Language Model. *ArXiv preprint* abs/2303.03378.

[98] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 736–740.

[99] Chenpeng Du, Yiwei Guo, Xie Chen, and Kai Yu. 2022. VQTTS: High-fidelity text-to-speech synthesis with self-supervised VQ acoustic feature. *arXiv preprint arXiv:2204.00768* (2022).

[100] Yexing Du, Ziyang Ma, Yifan Yang, Keqi Deng, Xie Chen, Bo Yang, Yang Xiang, Ming Liu, and Bing Qin. 2024. CoT-ST: Enhancing LLM-based Speech Translation with Multimodal Chain-of-Thought. *arXiv preprint arXiv:2409.19510* (2024).

[101] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407* (2024).

[102] Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, et al. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673* (2023).

[103] Pierre Duhamel and Martin Vetterli. 1990. Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing* 19, 4 (1990), 259–299.

[104] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High Fidelity Neural Audio Compression. *arXiv preprint arXiv:2210.13438* (2022).

[105] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. arXiv:2410.00037 [eess.AS] https://arxiv.org/abs/2410.00037

[106] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. arXiv:2410.00037 [eess.AS] https://arxiv.org/abs/2410.00037

[107] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. 2024. Scalable Pre-training of Large Autoregressive Image Models. arXiv:2401.08541 [cs.CV]

[108] Alaaeldin El-Nouby, MatthewJ. Muckley, Karen Ullrich, Ivan Laptev, Jakob Verbeek, and Hervé Jégou. 2022. Image Compression with Product Quantized Masked Image Modeling. (Dec 2022).

[109] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[110] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. CLAP: Learning Audio Concepts From Natural Language Supervision. *arXiv preprint arXiv: 2206.04769* (2022).

[111] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 336–340.

[112] Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming Transformers for High-Resolution Image Synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 12868–12878. https://api.semanticscholar.org/CorpusID:229297973

[113] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. 2024. Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens. arXiv:2410.13863 [cs.CV] https://arxiv.org/abs/2410.13863

[114] Jiarui Fang and Shangchun Zhao. 2024. USP: A Unified Sequence Parallelism Approach for Long Context Generative AI. arXiv:2405.07719 [cs.LG] https://arxiv.org/abs/2405.07719

[115] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666* (2024).

[116] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=fO31YAyNbI

[117] Weiqi Feng, Yangrui Chen, Shaoyu Wang, Yanghua Peng, Haibin Lin, and Minlan Yu. 2024. Optimus: Accelerating Large-Scale Multi-Modal LLM Training by Bubble Exploitation. arXiv:2408.03505 [cs.CL] https://arxiv.org/abs/2408.03505

[118] Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. 2022. Language models can learn complex molecular distributions. *Nature Communications* 13, 1 (2022), 3293.

[119] Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. 2022. Language models can learn complex molecular distributions. *Nature Communications* 13, 1 (June 2022). https://doi.org/10.1038/s41467-022-30839-x

[120] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394* (2023).

[121] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. MME-Survey: A Comprehensive Survey on Evaluation of Multimodal LLMs. arXiv:2411.15296 [cs.CV] https://arxiv.org/abs/2411.15296

[122] Sadaoki Furui. 1986. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 11. IEEE, 1991–1994.

[123] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108* (2023).

[124] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. https://doi.org/10.48550/ARXIV.2203.13131

[125] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[126] Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, and Benyou Wang. 2023. MLLM-Bench, Evaluating Multi-modal LLMs using GPT-4V. arXiv:2311.13951 [cs.CL]

[127] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041* (2023).

[128] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making LLaMA SEE and Draw with SEED Tokenizer. *arXiv preprint arXiv:2310.01218* (2023).

[129] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[130] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions. arXiv:2404.07214 [cs.CV] https://arxiv.org/abs/2404.07214

[131] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. 2023. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. arXiv:2310.11513 [cs.CV] https://arxiv.org/abs/2310.11513

[132] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. *arXiv preprint arXiv: 2305.05665* (2023).

[133] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[134] Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: Audio Spectrogram Transformer. *INTERSPEECH* (2021). https://doi.org/10.21437/interspeech.2021-698

[135] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790* (2023).

[136] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 6325–6334.

[137] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[138] R. Gray. 1984. Vector quantization. *IEEE ASSP Magazine* 1, 2 (1984), 4–29. https://doi.org/10.1109/MASSP.1984.1162229

[139] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing* 32, 2 (1984), 236–243.

[140] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. arXiv:2310.14566 [cs.CV]

[141] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).

[142] Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394* (2023).

[143] Hao-Han Guo, Kun Liu, Fei-Yu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283* (2024).

[144] Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2440–2452.

[145] Tanmay Gupta and Aniruddha Kembhavi. 2022. Visual Programming: Compositional visual reasoning without training. arXiv:2211.11559 [cs.CV]

[146] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. AudioCLIP: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 976–980.

[147] Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. C4Corpus: Multilingual Web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 914–922.

[148] Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. 2024. VALL-E R: Robust and Efficient Zero-Shot Text-to-Speech Synthesis via Monotonic Alignment. *arXiv preprint arXiv:2406.07855* (2024).

[149] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2024. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. arXiv:2412.04431 [cs.CV] https://arxiv.org/abs/2412.04431

[150] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language Models are General-Purpose Interfaces. arXiv:2206.06336 [cs.CL] https://arxiv.org/abs/2206.06336

[151] William Harvey and Frank Wood. 2023. Visual chain-of-thought diffusion models. *arXiv preprint arXiv:2303.16187* (2023).

[152] Jakob Drachmann Havtorn, Amelie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. 2023. MSViT: Dynamic Mixed-Scale Tokenization for Vision Transformers. *arXiv preprint arXiv: 2307.02321* (2023).

[153] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV] https://arxiv.org/abs/2111.06377

[154] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv: 1512.03385* (2015).

[155] Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. 2020. Query-Key Normalization for Transformers. arXiv:2010.04245 [cs.CL] https://arxiv.org/abs/2010.04245

[156] Amir Hertz, Ron Mokady, Jay M. Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *ArXiv* abs/2208.01626 (2022). https://api.semanticscholar.org/CorpusID:251252882

[157] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500 [cs.LG] https://arxiv.org/abs/1706.08500

[158] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

[159] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556 [cs.CL] https://arxiv.org/abs/2203.15556

[160] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[161] Ke Hu, Zhehuai Chen, Chao-Han Huck Yang, Piotr Żelasko, Oleksii Hrinchuk, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Chain-of-Thought Prompting for Speech Translation. *arXiv preprint arXiv:2409.11538* (2024).

[162] Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656* (2024).

[163] Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang, and Ming Zhang. 2024. MMEvalPro: Calibrating Multimodal Benchmarks Towards Trustworthy and Efficient Evaluation. arXiv:2407.00468 [cs.CV] https://arxiv.org/abs/2407.00468

[164] Jiaxing Huang and Jingyi Zhang. 2024. A Survey on Evaluation of Multimodal Large Language Models. arXiv:2408.15769 [cs.CV] https://arxiv.org/abs/2408.15769

[165] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. *Advances in Neural Information Processing Systems* 35 (2022), 28708–28720.

[166] Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. 2023. Make-a-voice: Unified voice synthesis with discrete representation. *arXiv preprint arXiv:2305.19269* (2023).

[167] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2023. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463* (2023).

[168] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023. VBench: Comprehensive Benchmark Suite for Video Generative Models. arXiv:2311.17982 [cs.CV]

[169] Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin CUI, and Wenming Yang. 2024. Protein-Ligand Interaction Prior for Binding-aware 3D Molecule Diffusion Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=qH9nrMNTIW

[170] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. HQ-Edit: A High-Quality Dataset for Instruction-based Image Editing. *arXiv preprint arXiv:2404.09990* (2024).

[171] Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M^2 UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255* (2023).

[172] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773 If you use this software, please cite it as below..

[173] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017* (2022).

[174] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models. arXiv:2309.14509 [cs.LG] https://arxiv.org/abs/2309.14509

[175] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv preprint arXiv: 2107.14795* (2021).

[176] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General Perception with Iterative Attention. *arXiv preprint arXiv: 2103.03206* (2021).

[177] Teerapat Jenrungrot, Michael Chinen, W Bastiaan Kleijn, Jan Skoglund, Zalán Borsos, Neil Zeghidour, and Marco Tagliasacchi. 2023. Lmcodec: A low bitrate speech codec with causal transformer models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[178] Shengpeng Ji, Ziyue Jiang, Xize Cheng, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, et al. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532* (2024).

[179] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[180] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). 4904–4916.

[181] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2024. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* 36 (2024).

[182] Kai Jiang and Jiaxing Huang. 2024. A Survey on Vision Autoregressive Model. arXiv:2411.08666 [cs.CV] https://arxiv.org/abs/2411.08666

[183] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2023. VIMA: General Robot Manipulation with Multimodal Prompts. arXiv:2210.03094 [cs.RO]

[184] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H. Chen, and Andrew Y. Ng. 2024. Many-Shot In-Context Learning in Multimodal Foundation Models. arXiv:2405.09798 [cs.LG] https://arxiv.org/abs/2405.09798

[185] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Efficient Multimodal Large Language Models: A Survey. arXiv:2405.10739 [cs.CV] https://arxiv.org/abs/2405.10739

[186] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. 2024. Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization. *arXiv preprint arXiv:2402.03161* (2024).

[187] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Yadong Mu, et al. 2023. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. *arXiv preprint arXiv:2309.04669* (2023).

[188] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv:1603.08155 [cs.CV] https://arxiv.org/abs/1603.08155

[189] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100* (2024).

[190] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7669–7673.

[191] Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10991–10995.

[192] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv preprint* abs/2001.08361 (2020).

[193] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics* 11 (2023), 1703–1718.

[194] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). 5583–5594.

[195] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114

[196] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *ICCV* (2023).

[197] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).

[198] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating Images with Multimodal Language Models. *NeurIPS* (2023).

[199] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David Ross, Grant Schindler, Mikhail Sirotenko, Kihyuk Sohn, Krishna Somandepalli, Huisheng Wang, Jimmy Yan, Ming-Hsuan Yang, Xuan Yang, Bryan Seybold, and Lu Jiang. 2023. VideoPoet: A Large Language Model for Zero-Shot Video Generation. arXiv:2312.14125 [cs.CV]

[200] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems* 33 (2020), 17022–17033.

[201] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831* (2024).

[202] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352* (2022).

[203] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123 (2017), 32–73.

[204] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2024. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems* 36 (2024).

[205] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180 [cs.LG] https://arxiv.org/abs/2309.06180

[206] Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. 2024. BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100K hours of data. *arXiv preprint arXiv:2402.08093* (2024).

[207] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics* 9 (2021), 1336–1354.

[208] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics* 9 (2021), 1336–1354.

[209] Luca Lanzendörfer, Florian Grötschla, Emil Funke, and Roger Wattenhofer. 2024. DISCO-10M: A Large-Scale Music Dataset. *Advances in Neural Information Processing Systems* 36 (2024).

[210] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. arXiv:1512.09300 [cs.LG] https://arxiv.org/abs/1512.09300

[211] Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of Large Audio Models: A Survey and Outlook. *arXiv preprint arXiv:2308.12792* (2023).

[212] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* 36 (2024).

[213] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527* (2023).

[214] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. arXiv:2306.16527 [cs.IR]

[215] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. 2009. Evaluation of algorithms using games: The case of music tagging.. In *ISMIR*. Citeseer, 387–392.

[216] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive Image Generation using Residual Quantization. arXiv:2203.01941 [cs.CV]

[217] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive Image Generation using Residual Quantization. *Cornell University - arXiv,Cornell University - arXiv* (Mar 2022).

[218] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2022. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. *arXiv preprint arXiv: 2210.03347* (2022).

[219] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. 2021. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10274–10284.

[220] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7331–7341.

[221] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023. SEED-Bench-2: Benchmarking Multimodal Large Language Models. arXiv:2311.17092 [cs.CV]

[222] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. arXiv:2307.16125 [cs.CL]

[223] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. MIMIC-IT: Multi-Modal In-Context Instruction Tuning. (2023). arXiv:2306.05425 [cs.CV]

[224] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. arXiv:2309.10020 [cs.CV] https://arxiv.org/abs/2309.10020

[225] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2024).

[226] Dongxu Li, Junnan Li, and Steven C. H. Hoi. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. arXiv:2305.14720 [cs.CV] https://arxiv.org/abs/2305.14720

[227] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint* abs/2301.12597 (2023).

[228] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

[229] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. arXiv:2311.17005 [cs.CV]

[230] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023. Silkie: Preference Distillation for Large Visual Language Models. (2023).

[231] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. M^{3}IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *ArXiv preprint* abs/2306.04387 (2023).

[232] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. 2023. M3DBench: Let's Instruct Large Models with Multi-modal 3D Prompts. *arXiv preprint arXiv:2312.10763* (2023).

[233] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024. Autoregressive Image Generation without Vector Quantization. arXiv:2406.11838

[234] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024. Autoregressive Image Generation without Vector Quantization. arXiv:2406.11838 [cs.CV] https://arxiv.org/abs/2406.11838

[235] Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. Yodas: Youtube-Oriented Dataset for Audio and Speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1–8.

[236] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. arXiv:2205.14217 [cs.CL] https://arxiv.org/abs/2205.14217

[237] Yazhe Li, Jorg Bornschein, and Ting Chen. 2024. Denoising Autoregressive Representation Learning. arXiv:2403.05196 [cs.LG]

[238] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *ArXiv preprint* abs/2305.10355 (2023).

[239] Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. 2023. Lmeye: An interactive perception network for large language models. *arXiv preprint arXiv:2305.03701* (2023).

[240] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring Plain Vision Transformer Backbones for Object Detection. arXiv:2203.16527 [cs.CV] https://arxiv.org/abs/2203.16527

[241] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. 2023. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107* (2023).

[242] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. 2023. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. *arXiv preprint arXiv: 2306.00107* (2023).

[243] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, et al. 2022. Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. *arXiv preprint arXiv:2212.02508* (2022).

[244] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253* (2023).

[245] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models. arXiv:2403.18814 [cs.CV] https://arxiv.org/abs/2403.18814

[246] Haokun Lin, Haoli Bai, Zhili Liu, Lu Hou, Muyi Sun, Linqi Song, Ying Wei, and Zhenan Sun. 2024. MoPE-CLIP: Structured Pruning for Efficient Vision-Language Models with Module-wise Pruning Error Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27370–27380.

[247] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17949–17958.

[248] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[249] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[250] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

[251] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. arXiv:2311.10774 [cs.CL]
[252] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *arXiv preprint arXiv: 2301.12503* (2023).
[253] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning.
[254] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *ArXiv preprint* abs/2304.08485 (2023).
[255] Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. 2024. SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound. *arXiv preprint arXiv:2405.00233* (2024).
[256] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253 [cs.CV] https://arxiv.org/abs/2402.00253
[257] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024. World Model on Million-Length Video and Language with RingAttention. *arXiv preprint* (2024).
[258] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024. AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining. arXiv:2308.05734 [cs.SD] https://arxiv.org/abs/2308.05734
[259] Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context. arXiv:2310.01889 [cs.CL] https://arxiv.org/abs/2310.01889
[260] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2023. Prismer: A Vision-Language Model with Multi-Task Experts. *arXiv preprint arXiv: 2303.02506* (2023).
[261] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. MMBench: Is Your Multi-modal Model an All-around Player? arXiv:2307.06281 [cs.CV]
[262] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. TempCompass: Do Video LLMs Really Understand Videos? arXiv:2403.00476 [cs.CV]
[263] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. TextMonkey: An OCR-Free Large Multimodal Model for Understanding Document. *arXiv preprint arXiv: 2403.04473* (2024).
[264] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. arXiv:2111.09883 [cs.CV]
[265] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV]
[266] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. arXiv:2201.03545 [cs.CV] https://arxiv.org/abs/2201.03545
[267] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. 2021. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems* 34 (2021), 28092–28103.
[268] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525 [cs.AI] https://arxiv.org/abs/2403.05525
[269] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. *arXiv preprint arXiv: 2312.17172* (2023).
[270] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. arXiv:2312.17172 [cs.CV]
[271] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. arXiv:2206.08916 [cs.CV]
[272] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv:2310.02255 [cs.CV]
[273] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
[274] Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. 2023. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110* (2023).
[275] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2024. Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training.

arXiv:2410.08202 [cs.CV] https://arxiv.org/abs/2410.08202

[276] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[277] Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, et al. 2024. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232* (2024).

[278] Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. 2024. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840* (2024).

[279] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093* (2023).

[280] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).

[281] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* 41, 8 (2023), 1099–1106.

[282] Brielen Madureira. 2021. Flamingos and Hedgehogs in the Croquet-Ground: Teaching Evaluation of NLP Systems for Undergraduate Students. In *Proceedings of the Fifth Workshop on Teaching NLP*. 87–91.

[283] Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. VoxtLM: Unified Decoder-Only Models for Consolidating Speech Recognition, Synthesis and Speech, Text Continuation Tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 13326–13330.

[284] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. 2023. Token Pooling in Vision Transformers for Image Classification. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*. IEEE, 12–21. https://doi.org/10.1109/WACV56688.2023.00010

[285] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[286] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395* (2023).

[287] Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Large Language Model Can Transcribe Speech in Multi-Talker Scenarios with Versatile Instructions. *arXiv preprint arXiv:2409.08596* (2024).

[288] Lingwei Meng, Jiawen Kang, Yuejiao Wang, Zengrui Jin, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Empowering Whisper as a Joint Multi-Talker and Target-Talker Speech Recognition System. In *Interspeech 2024*. 4653–4657. https://doi.org/10.21437/Interspeech.2024-971

[289] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al. 2024. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551* (2024).

[290] Gaurav Menghani. 2023. Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *Comput. Surveys* 55, 12 (March 2023), 1–37. https://doi.org/10.1145/3578938

[291] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite Scalar Quantization: VQ-VAE Made Simple. (Sep 2023).

[292] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2630–2640.

[293] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14420–14431.

[294] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. IMU2CLIP: Multimodal Contrastive Learning for IMU Motion Sensors from Egocentric Videos and Text. *arXiv preprint arXiv: 2210.14395* (2022).

[295] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. 2023. AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model. *arXiv preprint arXiv: 2309.16058* (2023).

[296] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. arXiv:2307.15189 [cs.CV]

[297] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*. Springer, 407–426.

[298] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. arXiv:2104.04473 [cs.CL] https://arxiv.org/abs/2104.04473

[299] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems* 35 (2022), 21455–21469.

[300] OpenAI. 2022. Introducing ChatGPT. (2022).

[301] OpenAI. 2023. GPT-4V(ision) System Card. (2023).

[302] OpenAI. 2024. hello-gpt-4o. https://openai.com/index/hello-gpt-4o/

[303] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* 24 (2011).

[304] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[305] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[306] Gilles Pagès. 2015. Introduction to vector quantization and its applications for numerics. *Esaim: Proceedings* 48 (2015), 29–79. https://api.semanticscholar.org/CorpusID:56105648

[307] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-G: Generating Images in Context with Multimodal Large Language Models. arXiv:2310.02992 [cs.CV] https://arxiv.org/abs/2310.02992

[308] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[309] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. 2024. aMUSEd: An Open MUSE Reproduction. arXiv:2401.01808 [cs.CV]

[310] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. arXiv:2406.17557 [cs.CL]

[311] Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. *arXiv preprint arXiv:2403.16973* (2024).

[312] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *arXiv preprint arXiv: 2208.06366* (2022).

[313] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824* (2023).

[314] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. 2023. Combined scaling for zero-shot transfer learning. *Neurocomputing* 555 (2023), 126658.

[315] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167* (2023).

[316] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV] https://arxiv.org/abs/2307.01952

[317] Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967* (2022).

[318] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. 2024. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236* (2024).

[319] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv:2307.16789 [cs.AI]

[320] Xingwei Qu, Yuelin Bai, Yinghao Ma, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, et al. 2024. MuPT: A Generative Symbolic Music Pretrained Transformer. *arXiv preprint arXiv:2404.06393* (2024).

[321] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). 8748–8763.

[322] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[323] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[324] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. 2024. Humanoid Locomotion as Next Token Prediction. *arXiv preprint arXiv:2402.19469* (2024).

[325] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

[326] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[327] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]

[328] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV] https://arxiv.org/abs/2102.12092

[329] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. arXiv:1907.01341 [cs.CV] https://arxiv.org/abs/1907.01341

[330] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).

[331] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:173990382

[332] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. 2023. TESTA: Temporal-spatial token aggregation for long-form video-language understanding. *arXiv preprint arXiv:2310.19060* (2023).

[333] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14313–14323.

[334] Jonathan Rohleder. 2024. A variational approach to the hot spots conjecture. arXiv:2404.01890 [math.SP] https://arxiv.org/abs/2404.01890

[335] Jason Tyler Rolfe. 2017. Discrete Variational Autoencoders. arXiv:1609.02200 [stat.ML]

[336] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr52688.2022.01042

[337] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[338] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[339] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925* (2023).

[340] Jeffrey A Ruffolo and Ali Madani. 2024. Designing proteins with language models. *Nature Biotechnology* 42, 2 (2024), 200–202.

[341] Jeffrey A. Ruffolo and Ali Madani. 2024. Designing proteins with language models. *Nature Biotechnology* 42 (2024), 200–202. https://api.semanticscholar.org/CorpusID:267682839

[342] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[343] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. arXiv:1904.01201 [cs.CV] https://arxiv.org/abs/1904.01201

[344] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[345] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *ArXiv preprint* abs/2111.02114 (2021).

[346] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Unleashing Chain-of-Thought Reasoning in Multi-Modal Language Models. *arXiv preprint arXiv:2403.16999* (2024).

[347] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[348] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[349] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv:2303.17580 [cs.CL]

[350] Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, and Nenghai Yu. 2024. Towards More Unified In-context Visual Understanding. arXiv:2312.02520 [cs.CV]

[351] Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2023. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *arXiv preprint arXiv:2310.00647* (2023).

[352] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. arXiv:2402.00159 [cs.CL] https://arxiv.org/abs/2402.00159

[353] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. arXiv:2307.16449 [cs.CV] https://arxiv.org/abs/2307.16449

[354] Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen. 2024. ELLA-V: Stable neural codec language modeling with alignment-guided sequence reordering. *arXiv preprint arXiv:2401.07333* (2024).

[355] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2443–2449.

[356] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. arXiv:2406.06525 [cs.CV] https://arxiv.org/abs/2406.06525

[357] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023. Generative Multimodal Models are In-Context Learners. *arXiv preprint arXiv:2312.13286* (2023).

[358] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative Multimodal Models are In-Context Learners. arXiv:2312.13286 [cs.CV] https://arxiv.org/abs/2312.13286

[359] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative Multimodal Models are In-Context Learners. arXiv:2312.13286 [cs.CV]

[360] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv: 2303.15389* (2023).

[361] Qingyun Sun and Zhen Guo. 2024. Scaling Law Hypothesis for Multimodal Model. arXiv:2409.06754 [cs.LG] https://arxiv.org/abs/2409.06754

[362] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative Pretraining in Multimodality. arXiv:2307.05222 [cs.CV]

[363] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Emu: Generative Pretraining in Multimodality. arXiv:2307.05222 [cs.CV]
[364] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. *ArXiv preprint* abs/2309.14525 (2023).
[365] Yan Tai, Weichen Fan, Zhao Zhang, Feng Zhu, Rui Zhao, and Ziwei Liu. 2023. Link-Context Learning for Multimodal LLMs. arXiv:2308.07891 [cs.CV]
[366] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289* (2023).
[367] Zecheng Tang, Chenfei Wu, Zekai Zhang, Mingheng Ni, Shengming Yin, Yu Liu, Zhengyuan Yang, Lijuan Wang, Zicheng Liu, Juntao Li, and Nan Duan. 2024. StrokeNUWA: Tokenizing Strokes for Vector Graphic Synthesis. *arXiv preprint arXiv: 2401.17093* (2024).
[368] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023. CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation. arXiv:2311.18775 [cs.CV]
[369] Chameleon Team. 2024. Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv:2405.09818 [cs.CL]
[370] Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv: 2312.11805* (2023).
[371] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
[372] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.
[373] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. arXiv:2404.02905 [cs.CV] https://arxiv.org/abs/2404.02905
[374] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. arXiv:2404.02905 [cs.CV]
[375] Eric Todd, Millicent Li, Arnab Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function Vectors in Large Language Models. In *International Conference on Learning Representations*. ICLR.
[376] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. arXiv:2406.16860 [cs.CV] https://arxiv.org/abs/2406.16860
[377] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.
[378] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint* abs/2302.13971 (2023).
[379] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint* abs/2307.09288 (2023).
[380] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. 2019. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv: 1906.06423* (2019).
[381] Michael Tschannen, Cian Eastwood, and Fabian Mentzer. 2024. GIVT: Generative Infinite-Vocabulary Transformers. arXiv:2312.02116 [cs.CV]
[382] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. arXiv:2106.13884 [cs.CV]
[383] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* 12 (2016).
[384] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).
[385] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. *ArXiv* abs/1711.00937 (2017). https://api.semanticscholar.org/CorpusID:20282961
[386] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.

[387] Shubham Vatsal and Harsh Dubey. 2024. A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. arXiv:2407.12994 [cs.CL] https://arxiv.org/abs/2407.12994

[388] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable Length Video Generation From Open Domain Textual Description. arXiv:2210.02399 [cs.CV]

[389] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2023. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908* (2023).

[390] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient Large Language Models: A Survey. arXiv:2312.03863 [cs.CL] https://arxiv.org/abs/2312.03863

[391] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

[392] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. 2023. SAM-CLIP: Merging Vision Foundation Models towards Semantic and Spatial Understanding. *arXiv preprint arXiv: 2310.15308* (2023).

[393] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574* (2023).

[394] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. arXiv:2402.14804 [cs.CV]

[395] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191 [cs.CV] https://arxiv.org/abs/2409.12191

[396] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *CoRR* abs/2202.03052 (2022).

[397] Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107* (2023).

[398] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. 2023. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907* (2023).

[399] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2023. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. arXiv:2212.02499 [cs.CV]

[400] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. 2024. Emu3: Next-Token Prediction is All You Need. arXiv:2409.18869 [cs.CV] https://arxiv.org/abs/2409.18869

[401] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. 2024. Emu3: Next-Token Prediction is All You Need. arXiv:2409.18869 [cs.CV] https://arxiv.org/abs/2409.18869

[402] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning. arXiv:2401.06805 [cs.CL] https://arxiv.org/abs/2401.06805

[403] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942* (2023).

[404] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2024. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355* (2024).

[405] Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. 2024. MIO: A Foundation Model on Multimodal Tokens. *arXiv preprint arXiv: 2409.17692* (2024).

[406] Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2024. MaskBit: Embedding-free Image Generation via Bit Tokens. arXiv:2409.16211 [cs.CV] https://arxiv.org/abs/2409.16211

[407] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).

[408] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[409] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. 2022. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. arXiv:2210.10716 [cs.CV]

[410] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2024. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. arXiv:2410.13848 [cs.CV] https://arxiv.org/abs/2410.13848

[411] Haibin Wu, Kai-Wei Chang, Yuan-Kuei Wu, and Hung-yi Lee. 2023. Speechgen: Unlocking the generative power of speech language models with prompts. *arXiv preprint arXiv:2306.02207* (2023).

[412] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4563–4567.

[413] Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1–8.

[414] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. 2024. Towards A Better Metric for Text-to-Video Generation. arXiv:2401.07781 [cs.CV]

[415] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. 2023. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21970–21980.

[416] Penghao Wu and Saining Xie. 2024. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13084–13094.

[417] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv: 2309.05519* (2023). https://arxiv.org/abs/2309.05519v2

[418] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[419] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. 2024. Number it: Temporal Grounding Videos like Flipping Manga. *arXiv preprint arXiv:2411.10332* (2024).

[420] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2024. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation. arXiv:2409.04429 [cs.CV] https://arxiv.org/abs/2409.04429

[421] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large Multimodal Agents: A Survey. arXiv:2402.15116 [cs.CV] https://arxiv.org/abs/2402.15116

[422] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. arXiv:2408.12528 [cs.CV] https://arxiv.org/abs/2408.12528

[423] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. arXiv:2408.12528 [cs.CV] https://arxiv.org/abs/2408.12528

[424] Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725* (2024).

[425] Yizhe Xiong, Hui Chen, Tianxiang Hao, Zijia Lin, Jungong Han, Yuesong Zhang, Guoxin Wang, Yongjun Bao, and Guiguang Ding. 2024. PYRA: Parallel Yielding Re-activation for Training-Inference Efficient Task Adaptation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX (Lecture Notes in Computer Science, Vol. 15067)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 455–473. https://doi.org/10.1007/978-3-031-72673-6_25

[426] Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. 2023. Youku-mPLUG: A 10 Million Large-scale Chinese Video-Language Dataset for Pre-training and

Benchmarks. *arXiv preprint arXiv:2306.04362* (2023).

[427] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 5288–5296.

[428] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[429] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, Qiyang Zhang, Zhenyan Lu, Li Zhang, Shangguang Wang, Yuanchun Li, Yunxin Liu, Xin Jin, and Xuanzhe Liu. 2024. A Survey of Resource-efficient LLM and Multimodal Foundation Models. arXiv:2401.08092 [cs.LG] https://arxiv.org/abs/2401.08092

[430] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. *arXiv preprint arXiv:2403.11703* (2024).

[431] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *arXiv preprint arXiv: 1912.13318* (2019).

[432] Zhiyang Xu, Trevor Ashby, Chao Feng, Rulin Shao, Ying Shen, Di Jin, Qifan Wang, and Lifu Huang. 2023. Vision-Flan:Scaling Visual Instruction Tuning. https://vision-flan.github.io/

[433] Zi'an Xu, Yin Dai, Fayu Liu, Weibing Chen, Yue Liu, Lifu Shi, Sheng Liu, and Yuhang Zhou. 2023. Swin MAE: Masked Autoencoders for Small Datasets. arXiv:2212.13805 [cs.CV]

[434] Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning. *ArXiv preprint* abs/2212.10773 (2022).

[435] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5036–5045.

[436] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 483–498.

[437] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. VideoGPT: Video Generation using VQ-VAE and Transformers. arXiv:2104.10157 [cs.CV]

[438] Chenyu Yang, Xizhou Zhu, Jinguo Zhu, Weijie Su, Junjie Wang, Xuan Dong, Wenhai Wang, Lewei Lu, Bin Li, Jie Zhou, Yu Qiao, and Jifeng Dai. 2024. Vision Model Pre-training on Interleaved Image-Text Data via Latent Compression Learning. arXiv:2406.07543 [cs.CV] https://arxiv.org/abs/2406.07543

[439] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765* (2023).

[440] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. 2023. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704* (2023).

[441] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. 2023. Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model. *arXiv preprint arXiv:2311.13231* (2023).

[442] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2024. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv:2209.00796 [cs.LG] https://arxiv.org/abs/2209.00796

[443] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems* 36 (2024).

[444] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2024. Exploring Diverse In-Context Configurations for Image Captioning. arXiv:2305.14800 [cs.CV]

[445] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* (2021).

[446] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

[447] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Mingshi Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Feiyan Huang. 2023. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. *Conference on Empirical Methods in Natural Language Processing* (2023). https://doi.org/10.48550/arXiv.2310.05126

[448] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. arXiv:2306.13549 [cs.CV] https://arxiv.org/abs/2306.13549

[449] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. 2024. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems* 36 (2024).

[450] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2022. Vector-quantized Image Modeling with Improved VQGAN. arXiv:2110.04627 [cs.CV]

[451] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=Ee277P3AYC

[452] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[453] Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, and Joyce Chai. 2023. Efficient In-Context Learning in Vision-Language Models for Egocentric Videos. arXiv:2311.17041 [cs.CV]

[454] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. 2022. MAGVIT: Masked Generative Video Transformer. *arXiv preprint arXiv: 2212.05199* (2022).

[455] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A. Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, Kevin P. Murphy, Alexander G. Hauptmann, and Lu Jiang. 2023. SPAE: Semantic Pyramid AutoEncoder for Multimodal Generation with Frozen LLMs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/a526cc8f6ffb74bedb6ff313e3fdb450-Abstract-Conference.html

[456] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2024. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. arXiv:2310.05737 [cs.CV]

[457] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2023. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. arXiv:2310.05737 [cs.CV]

[458] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. 2023. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning. arXiv:2309.02591 [cs.LG]

[459] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. 2024. Randomized Autoregressive Visual Generation. arXiv:2411.00776 [cs.CV] https://arxiv.org/abs/2411.00776

[460] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. 2023. CapsFusion: Rethinking Image-Text Data at Scale. *arXiv preprint arXiv:2310.20550* (2023).

[461] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2024. An Image is Worth 32 Tokens for Reconstruction and Generation. arXiv:2406.07550 [cs.CV] https://arxiv.org/abs/2406.07550

[462] Shixing Yu, Tianlong Chen, Jiayi Shen, Huan Yuan, Jianchao Tan, Sen Yang, Ji Liu, and Zhangyang Wang. 2022. Unified Visual Transformer Compression. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=9jsZiUgkCZP

[463] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. *arxiv* (2023).

[464] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).

[465] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*. 558–567.

[466] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo

Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv:2311.16502 [cs.CL]

[467] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 495–507.

[468] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (Jan 2022), 495–507. https://doi.org/10.1109/taslp.2021.3129994

[469] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16375–16387.

[470] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882* (2019).

[471] Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs Perform Text-to-Image In-Context Learning? arXiv:2402.01293 [cs.LG]

[472] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12104–12113.

[473] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343 [cs.CV] https://arxiv.org/abs/2303.15343

[474] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. *ArXiv* abs/2402.12226 (2024). https://api.semanticscholar.org/CorpusID:267750101

[475] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6182–6186.

[476] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000* (2023).

[477] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. arXiv:2401.13601 [cs.CL] https://arxiv.org/abs/2401.13601

[478] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, Wenhu Chen, and Jie Fu. 2024. CMMMU: A Chinese Massive Multi-discipline Multimodal Understanding Benchmark. arXiv:2401.11944 [cs.CL]

[479] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-Language Models for Vision Tasks: A Survey. arXiv:2304.00685 [cs.CV] https://arxiv.org/abs/2304.00685

[480] Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024. Benchmarking Large Multimodal Models against Common Corruptions. arXiv:2401.11943 [cs.LG]

[481] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2024. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* 36 (2024).

[482] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? arXiv:2403.14624 [cs.CV]

[483] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. 2023. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618* (2023).

[484] Wentao Zhang, Junliang Guo, Tianyu He, Li Zhao, Linli Xu, and Jiang Bian. 2024. Video In-context Learning. arXiv:2407.07356 [cs.CV] https://arxiv.org/abs/2407.07356

[485] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* (2023).

[486] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechTokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692* (2023).

[487] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *arXiv preprint arXiv: 2303.01037* (2023).

[488] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219 [cs.CL] https://arxiv.org/abs/2309.01219

[489] Yuhui Zhang, Brandon McKinzie, Zhe Gan, Vaishaal Shankar, and Alexander Toshev. 2023. Pre-trained Language Models Do Not Help Auto-regressive Text-to-Image Generation. arXiv:2311.16201 [cs.CV]

[490] Yuhui Zhang, Brandon McKinzie, Zhe Gan, Vaishaal Shankar, and Alexander Toshev. 2024. Pre-trained Language Models Do Not Help Auto-regressive Text-to-Image Generation. arXiv:2311.16201 [cs.CV] https://arxiv.org/abs/2311.16201

[491] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107* (2023).

[492] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2023. Universal Multimodal Representation for Language Understanding. *arXiv preprint arXiv: 2301.03344* (2023).

[493] Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, et al. 2024. Speechlm: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[494] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

[495] Zili Zhang, Yinmin Zhong, Ranchen Ming, Hanpeng Hu, Jianjian Sun, Zheng Ge, Yibo Zhu, and Xin Jin. 2024. DistTrain: Addressing Model and Data Heterogeneity with Disaggregated Training for Multimodal Large Language Models. arXiv:2408.04275 [cs.DC] https://arxiv.org/abs/2408.04275

[496] Bo Zhao, Boya Wu, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087* (2023).

[497] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. *ArXiv preprint* abs/2309.07915 (2023).

[498] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. arXiv:2407.05282 [cs.CV] https://arxiv.org/abs/2407.05282

[499] Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. 2024. Looking Beyond Text: Reducing Language bias in Large Vision-Language Models via Multimodal Dual-Attention and Soft-Image Guidance. arXiv:2411.14279 [cs.CV] https://arxiv.org/abs/2411.14279

[500] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. 2023. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474* (2023).

[501] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223

[502] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581* (2023).

[503] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *arXiv preprint arXiv:2305.16103* (2023).

[504] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. 2022. MoVQ: Modulating Quantized Vectors for High-Fidelity Image Generation. arXiv:2209.09002 [cs.CV] https://arxiv.org/abs/2209.09002

[505] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. DDCoT: Duty-Distinct Chain-of-Thought Prompting for Multimodal Reasoning in Language Models. arXiv:2310.16436 [cs.CV]

[506] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens.

[507] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. arXiv:2408.11039 [cs.AI] https://arxiv.org/abs/2408.11039

[508] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual In-Context Learning for Large Vision-Language Models. arXiv:2402.11574 [cs.CV]

[509] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *ArXiv preprint* abs/2304.10592 (2023).

[510] Hongyuan Zhu, Ye Niu, Di Fu, and Hao Wang. 2021. MusicBERT: A self-supervised learning of music representation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3955–3963.

[511] Mingjian Zhu, Kai Han, Yehui Tang, and Yunhe Wang. 2021. Visual Transformer Pruning. *CoRR* abs/2104.08500 (2021). arXiv:2104.08500 https://arxiv.org/abs/2104.08500

[512] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939* (2023).