Show-o2: Improved Native Unified Multimodal Models

Jinheng Xie[1] **Zhenheng Yang**[2] **Mike Zheng Shou**[1*]

[1] Show Lab, National University of Singapore [2] ByteDance

Abstract

This paper presents improved native unified multimodal models, *i.e.,* Show-o2, that leverage autoregressive modeling and flow matching. Built upon a 3D causal variational autoencoder space, unified visual representations are constructed through a dual-path of spatial (-temporal) fusion, enabling scalability across image and video modalities while ensuring effective multimodal understanding and generation. Based on a language model, autoregressive modeling and flow matching are natively applied to the language head and flow head, respectively, to facilitate text token prediction and image/video generation. A two-stage training recipe is designed to effectively learn and scale to larger models. The resulting Show-o2 demonstrates versatility in handling a wide range of multimodal understanding and generation tasks across diverse modalities, including text, images, and videos. Code and models are released at https://github.com/showlab/Show-o.

1 Introduction

Large language models (LLMs) [89, 107] have achieved unprecedented performance levels, fueled by extensive web-scale text resources, substantial computational power, and billions of parameters. In the multimodal domain, large multimodal models (LMMs) [6, 26, 50] and visual generative models [33, 76, 102], have also demonstrated exceptional capabilities in tasks such as general-purpose visual question answering and text-to-image/video generation. Given their success, unified multimodal models (UMMs) [86, 99, 104] have been investigated to unify multimodal understanding and generation within a single model or system. In addition to multimodal understanding capability, this line of approaches seeks to simultaneously cultivate visual generation ability in the model/system through pre-training, fine-tuning, or connecting tailored models.

Here, we provide a comparative analysis of selected UMMs in Table 1, focusing on two perspectives, including i) visual representations for understanding and generation and ii) the type of unified modeling. Generally, there are two approaches to incorporating visual representations for multimodal understanding and generation: i) a unified representation for both understanding and generation, as seen in works like Chameleon [86], Transfusion [117], and Show-o [104]; and ii) decoupled representations, utilizing CLIP [78] for multimodal understanding and variational autoencoder (VAE) for visual generation. To involve both multimodal understanding and generation capabilities, two primary methods have been explored: i) natively applying multimodal understanding and generation objectives within a single model and ii) tuning adapters to assemble tailored models. We refer the first type as *native unified multimodal models*, distinguishing it from the second type that assembles tailored models. These principles, combined with autoregressive or diffusion modeling or both, contribute to the development of unified multimodal models.

Compared to existing UMMs that primarily focus on text and image, our approach explores model designs that provide substantial potential and scalability in unifying text, image, and video modalities. An overview of our approach is presented in Fig. 1. Specifically, for visual inputs, we operate

* Corresponding Author

Table 1: Comparative analysis of selected unified multimodal models based on the type of visual representations and unified modeling for multimodal understanding and generation. In this context, **native und. & gen.** refers to the direct decoding of output sequences into texts, images, and videos, as opposed to serving as conditions for decoding using external pre-trained decoders like Stable Diffusion. * indicates the method adopts two distinct models for multimodal understanding and generation, respectively. Diff. means the diffusion modeling. *Please refer to the complete table in the appendix.*

Methods	Und. & Gen. Representation			Type of Unified Modeling		
	Unified	Decoupled	Support Video	Native Und. & Gen.	Assembling Tailored Models	Paradigm
Chameleon [86]	✓		✗	✓		AR
Transfusion [117]	✓		✗	✓		AR + Diff.
Show-o [104]	✓		✗	✓		AR + Diff.
VILA-U [101]	✓		✓	✓		AR
Emu3 [94]	✓		✓	✓		AR
Show-o2 (Ours)	✓		✓	✓		AR + Diff.
Janus-Series [23, 24, 70]		✓	✗	✓		AR (+Flow)
UnidFluid [34]		✓	✗	✓		AR + MAR
Mogao [59]		✓	✗	✓		AR + Diff.
Bagel [28]		✓	✓	✓		AR + Diff.
NExT-GPT [99]		✓	✓		✓	AR + Diff.
SEED-X [36]		✓	✗		✓	AR + Diff.
ILLUME [92]		✓	✗		✓	AR + Diff.
MetaMorph [88]		✓	✗		✓	AR + Diff.
TokenFlow* [77]	✓		✗		✓	AR
LlamaFusion [81]	✓		✗	✓		AR + Diff.

within the 3D causal VAE [90] space, which is capable of accommodating both images and videos. Recognizing the distinct feature dependencies between multimodal understanding and generation, we construct unified visual representations that simultaneously capture rich semantic information and low-level features with intrinsic structures and textual details from the visual latents. This is achieved through a dual-path mechanism consisting of semantic layers, a projector, and a spatial (-temporal) fusion process. As the fusion process occurs within the 3D causal VAE space, when it comes to videos, semantic and low-level features are temporally aligned and fused with full-frame video information.

Text embeddings and unified visual representations are structured into a sequence to go through a pre-trained language model and are modeled by a specific language head and flow head, respectively. Specifically, autoregressive modeling with causal attention is performed on the language head when dealing with text token prediction, and flow matching with full attention is applied to the flow head for image/video generation. Since the base language model lacks visual generation capabilities, we propose a two-stage training recipe to effectively learn such an ability while retaining the language knowledge, without requiring a massive text corpus. In the first stage, we mainly focus on pre-training the flow head for visual generation using (interleaved) text, image, and video data. In the second stage, the full model is fine-tuned with high-quality instruction and generation data.

Extensive experimental results have demonstrated that our model surpasses the existing methods in terms of most metrics across multimodal understanding and visual generation benchmarks. Collectively, the main contributions of this paper can be summarized as:

- We present an improved native unified multimodal model that seamlessly integrates autoregressive modeling and flow matching, enabling a wide range of multimodal understanding and generation across (interleaved) text, images, and videos.
- Based on the 3D causal VAE space, we construct unified visual representations scalable to both multimodal understanding and generation, image and video modalities by combining semantic and low-level features through a dual-path of spatial (-temporal) fusion mechanism.
- We design a two-stage training pipeline that effectively and efficiently learns unified multimodal models, retaining language knowledge and enabling effective scaling up to larger models, without requiring a massive text corpus.
- The proposed model demonstrates state-of-the-art performance on multimodal understanding and visual generation benchmarks, surpassing existing methods across various metrics.

2 Related Work

2.1 Large Multimodal Models

Building upon the advancements of large language models (LLMs) [89, 107], large multimodal models (LMMs) [6, 26, 50, 63] have showcased remarkable capabilities in general-purpose visual question answering. These approaches typically leverage pre-trained vision encoders to project visual features and align them within the embedding space of LLMs. Meanwhile, a growing number of encoder-free LMMs [30, 31, 104] aim to directly align raw visual features within the LLM embedding space. However, these encoder-free methods often fall behind models that utilize image-text-aligned visual features in terms of performance. Beyond model architecture, recent studies [18, 50, 87] have highlighted the critical role of high-quality instructional data in enhancing multimodal capabilities.

2.2 Visual Generative Models

Two prominent paradigms for visual generation, namely diffusion [8,17,61,75,76,80,98,102,103,113] and autoregressive modeling [20, 47, 53, 73, 83], have been extensively studied in image and video generation in recent years. Diffusion-based methods typically employ optimized architectures that integrate pre-trained text encoders with denoising networks. In contrast, autoregressive methods often utilize LLM-based architectures and are trained through next-token prediction. Recently, several studies [34, 55, 64] have explored hybrid approaches that combine diffusion and autoregressive modeling to further advance visual generation capabilities.

2.3 Unified Multimodal Models

Building on the success of large multimodal and visual generative models, pioneering unified multimodal models (UMMs) such as Chameleon [86], Show-o [104], and Transfusion [117] aim to integrate these capabilities into a single model through autoregressive or diffusion modeling or both. Further advancements [25, 45, 68, 82, 94, 101] have focused on optimizing the training pipeline and enhancing the semantics of discrete tokens, leading to improved performance. We refer to these approaches as *native unified multimodal models*, as they inherently combine multimodal understanding and generation objectives within a unified architecture.

An alternative and promising direction [32, 36, 67, 72, 85, 88] for unifying multimodal understanding and generation involves assembling off-the-shelf specialized LMMs and visual generative models by tuning adapters or learnable tokens. Representative works [36, 99] have demonstrated the promissing capabilities and intriguing properties of such assembled unified frameworks, highlighting their potential for further exploration.

3 Methodology

In this section, we introduce the overall framework (Section 3.1), which consists of two key components: i) the design of unified visual representations for multimodal understanding and generation, applicable to both images and videos, and ii) the native learning of multimodal understanding and generation capabilities. Subsequently, we present a two-stage training recipe (Section 3.2), which is designed to progressively learn and effectively scale up the unified multimodal model.

3.1 Overall Framework

Overall Architecture. An overview of our proposed unified model is depicted in Fig. 1. Given (interleaved) texts, images, or videos, a text tokenizer with an embedding layer and a 3D causal VAE encoder accordingly process them into continuous text embeddings and visual latent representations. Subsequently, the visual latent representations undergo a dual-path extraction of spatial (-temporal) fusion to create the unified visual representations. These representations are then structured into a sequence, which is fed into a language model equipped with language and flow heads to model the sequence via autoregressive modeling and flow matching accordingly. Finally, a text de-tokenizer in conjunction with a 3D causal VAE decoder is employed to decode the final output. Next, we will delve into the fundamental design principles behind the unified visual representation and flow head.

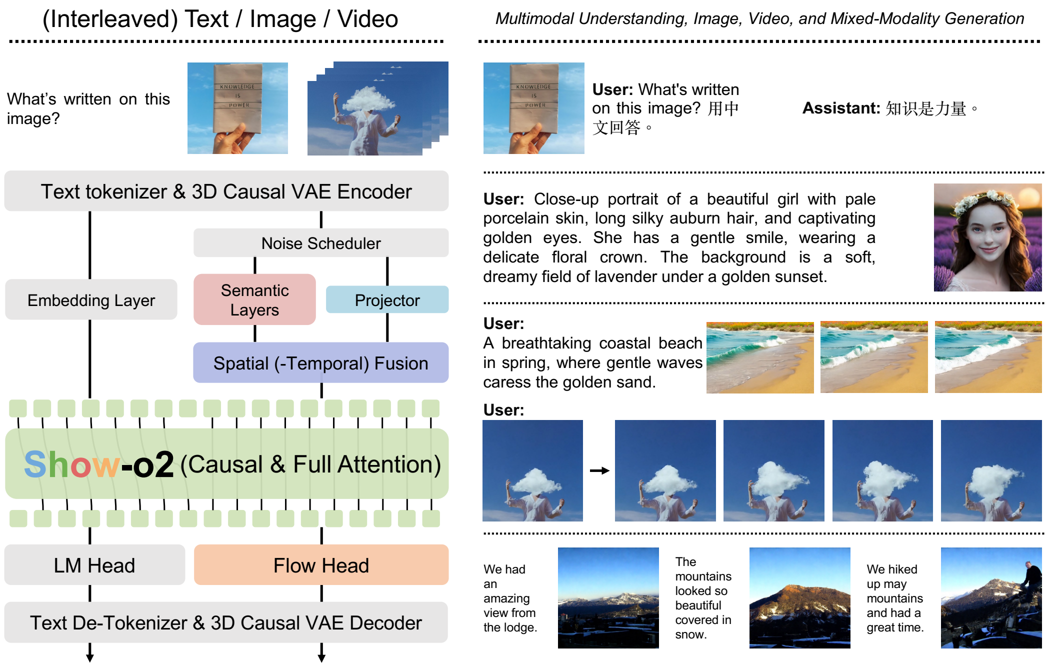

Figure 1: Our approach begins by encoding input texts, images, and videos into continuous embeddings and visual latents. The visual latents are processed through a dual-path extraction and spatial (-temporal) fusion mechanism to construct unified visual representations that are scalable for both multimodal understanding and generation, image and video modalities. These text embeddings and unified visual representations are then structured into a sequence for the base language model, equipped with dedicated heads. Specifically, text tokens are modeled autoregressively by a language head, while image and video latents are handled by a flow head using flow matching. We employ the omni-attention mechanism [104, 117] to enable causal attention along the sequence while maintaining full attention within the unified visual representations. This design empowers our model to effectively tackle tasks such as image/video understanding, generation, and mixed-modality generation.

Unified Visual Representation. To scalably support image and video modalities, we employ a 3D causal VAE encoder to extract image/video latents. As multimodal understanding and generation differ in feature dependency, we propose a dual-path architecture comprising semantic layers $\mathcal{S}(\cdot)$ to extract high-level representations of rich semantic contextual information and a projector $\mathcal{P}(\cdot)$ to retain complete low-level information from the extracted visual latents. Specifically, semantic layers $\mathcal{S}(\cdot)$ share the same vision transformer blocks of SigLIP [111] with a new 2×2 patch embedding layer. Given n visual latents $\mathbf{x}_t = \{x_i\}_{i=1}^{n}$ at a noise level:

$$\mathbf{x}_t = t \cdot \mathbf{x}_1 + (1 - t) \cdot \mathbf{x}_0, \tag{1}$$

where $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$ and $t \sim [0, 1]$, we load the pre-trained weights of SigLIP and pre-distill $\mathcal{S}(\cdot)$ as follows:

$$\mathcal{L}_{\text{distill}} = -\sum \log \text{sim}(\mathcal{S}(\mathbf{x}_t), \texttt{SigLIP}(\mathbf{X})), \tag{2}$$

where $\mathbf{X}$ is the input image, $\texttt{SigLIP}(\cdot)$ extracts the image patch features, and $\text{sim}(\cdot)$ indicates the cosine similarity calculator. In this way, semantic layers $\mathcal{S}(\cdot)$ can mimic extracting semantic features from both clean and noised visual latents $\mathbf{x}_t$. The projector $\mathcal{P}(\cdot)$ is simply composed of a 2D patch embedding layer. The extracted high- and low-level representations are spatially (and temporally when it comes to videos) fused by concatenating through the feature dimension and applying RMSNorm [112] with two MLP layers to get the unified visual representations $\mathbf{u}$:

$$\mathbf{u} = \texttt{STF}(\mathcal{S}(\mathbf{x}_t), \mathcal{P}(\mathbf{x}_t)), \tag{3}$$

where $\texttt{STF}$ indicates the spatial (-temporal) fusion mechanism. In addition, we prepend a time step t embedding to the unified visual representations for generative modeling. t is set as 1.0 to get time step embedding for the clean image.

We structure the text embeddings and unified visual representations into a sequence following a general interleaved image-text format below:

$$\text{[BOS]}\ \{\text{Text}\}\ \text{[BOI / BOV]}\ \{\text{Image / Video}\}\ \text{[EOI / EOV]}\ \{\text{Text}\} \cdots\ \text{[EOS]}.$$

The sequence format above is flexible and can be adapted to various input types. We adopt the omni-attention mechanism [104, 117] to let the sequence modeling be causal but with full attention within the unified visual representations.

Flow Head. Apart from the language head for text token prediction, we employ a flow head to predict the defined velocity $\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt}$ via flow matching [61, 65]. Specifically, the flow head simply consists of several transformer layers with time step modulation via the adaLN-Zero block, as seen in DiT [74].

During training, we natively apply next token prediction $\mathcal{L}_{\text{NTP}}$ to the language head and flow matching $\mathcal{L}_{\text{FM}}$ to the flow head for predicting velocity, respectively:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{NTP}} + \mathcal{L}_{\text{FM}}. \tag{4}$$

3.2 Training Recipe

Table 2: Trainable components and datasets in the training stages.

	Trainable Components	Datasets		
		# Image-Text	# Video-Text	# Interleaved Data
Stage-1	Projector Spatial (-Temporal) Fusion Flow Head	66M	WebVid [7] Pandas [21]	OmniCorpus [54]
Stage-2	Full Model (w/o VAE)	9M HQ Und. 16M HQ Gen.	OpenVid-1M [71] 1.5M Internal Data	VIST [42]

Existing UMMs, such as Show-o [104], Janus-Pro [23], Transfusion [117], Chameleon [86], and Emu3 [94], are typically trained from LLMs, LMMs, or from scratch. These approaches aim to cultivate visual generative modeling capabilities while preserving language modeling proficiency. However, this process often relies on web-scale, high-quality text corpora, which are prohibitively expensive to collect. Consequently, the lack of such resources can lead to a degradation in language knowledge and modeling performance. To address this challenge, we adopt a two-stage training recipe (as shown in Table 2) that effectively retains language knowledge while simultaneously developing visual generation capabilities, without requiring a massive text corpus.

Stage-1. Before the two-stage training, we have pre-distilled the semantic layers $\mathcal{S}(\cdot)$ (implementation details can be found in Section 4). The first stage only involves trainable components of the projector, spatial (-temporal) fusion, and flow head. In this stage, we train these components using autoregressive modeling and flow matching using around 66M image-text pairs and progressively add interleaved data and video-text pairs.

Stage-2. Subsequently, we tune the full model using 9M high-quality multimodal understanding instruction data and 16M high-quality visual generation data filtered from the 66M image-text pairs.

Scaling Up. After the training of the small-sized model with approximately 1.5B LLM parameters, we resume the pre-trained flow head for the larger model with 7B LLM parameters and introduce a lightweight MLP transformation to align the hidden size, allowing it to quickly adapt to the larger model and converge.

4 Experiments

4.1 Experimental Setup

Datasets. The curated approximately 66M image-text pairs consist of images with a resolution of at least 512 pixels in width and height. The images are filtered from CC12M [12], COYO [11], LAION-Aesthetic-12M* and AI synthetic data. The images are recaptioned by ShareGPT4-V [18] except for the synthetic data. The 9M high-quality multimodal understanding instruction data is curated from Densefusion-1M [56], and LLaVA-OneVision [50].

Implementation Details. The semantic layers $\mathcal{S}(\cdot)$ are pre-distilled from SigLIP-so400m-patch14-384* over 200K iterations, using a batch size of 512 and a cosine-scheduled learning rate of 2e-5.

* https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap
* https://huggingface.co/google/siglip-so400m-patch14-384

Table 3: Evaluation on multimodal understanding benchmarks. # Params. indicates the number of parameters of base LLM. * indicates the method uses two distinct models or sets of parameters for multimodal understanding and generation, respectively. Und. indicates "understanding".

Types	Models	# Params.	MME ↑ (p)	GQA↑	SEED ↑ (all)	MMB↑ (en-dev)	MMMU ↑ (val)	MMStar ↑	AI2D ↑
Und. Only	LLaVA-v1.5 [62]	7B	1510.7	62.0	58.6	64.3	-	-	-
	Qwen-VL-Chat [5]	7B	1487.6	57.5	58.2	60.6	-	-	57.7
Unify via Assembling Tailored Models	NExT-GPT [104]	13B	-	-	57.5	58.0	-	-	-
	SEED-X [36]	17B	1457.0	49.1	66.5	70.1	35.6	-	-
	MetaMorph [88]	8B	-	-	71.8	75.2	-	-	-
	TokenFlow-XL* [77]	14B	1551.1	62.5	72.6	76.8	43.2	-	75.9
	ILLUME [92]	7B	1445.3	-	72.9	75.1	38.2	-	71.4
Native Unified	Show-o [104]	1.3B	1097.2	58.0	51.5	-	27.4	-	-
	JanusFlow [70]	1.5B	1333.1	60.3	70.5	74.9	29.3	-	-
	SynerGen-VL [52]	2.4B	1381.0	-	-	53.7	34.2	-	-
	Janus-Pro [23]	1.5B	1444.0	59.3	68.3	75.5	36.3	-	-
	Show-o2 (Ours)	1.5B	1450.9	60.0	65.6	67.4	37.1	43.4	69.0
	Emu3 [94]	8B	-	60.3	68.2	58.5	31.6	-	70.0
	VILA-U [101]	7B	1401.8	60.8	59.0	-	-	-	-
	MUSE-VL [105]	7B	-	-	69.1	72.1	39.7	49.6	69.8
	Liquid [97]	8B	1448.0	61.1	-	-	-	-	-
	Janus-Pro [23]	7B	1567.1	62.0	72.1	79.2	41.0	-	-
	Mogao [59]	7B	1592.0	60.9	**74.6**	75.0	44.2	-	-
	Show-o2 (Ours)	7B	**1620.5**	**63.1**	69.8	**79.3**	**48.9**	**56.6**	**78.6**

During distillation, Eq. 1 is applied to the visual latents with only a probability of 0.3 in the last 20K iterations. The input image resolution of 3D causal VAE encoder with 2×2 patch embedding layer is set as 432×432 to get $729 = 27 \times 27$ visual latents, which matches the ones extracted by SigLIP. Once distilled, the semantic layers $\mathcal{S}(\cdot)$ are capable of extracting rich semantic features from both clean and noised visual latents. In statistics, the extracted features from clean visual latents by $\mathcal{S}(\cdot)$ have converged to an average cosine similarity of around 0.9 with those extracted by the original SigLIP on the curated 66M image-text pairs. We interpolate the position embeddings in the bicubic mode when involving other image/video resolutions.

Our models build upon two LLM variants, *i.e.,* Qwen2.5-1.5B-Instruct [107] and Qwen2.5-7B-Instruct [107], respectively. We adopt 3D causal VAE proposed in Wan2.1 [90] with $8\times$ and $4\times$ spatial and temporal compression, respectively. In stage 1, we first train the 1.5B variant for 150K iterations using AdamW optimizer with a constant learning rate of 0.0001 on the curated 66M image-text pairs in a resolution of 432×432. The context length of single image-text pairs is set as 1024. The total batch sizes for multimodal understanding and generation are 128 and 384, respectively. α in Eq. 4 is set as 0.2. For visual generation data, the caption is dropped with a probability of 0.1 to enable the classifier-free guidance. This training process roughly takes one and a half days using 64 H100 GPUs. Subsequently, we replace the generation data with 16M high-quality data (filtered from 66M image-text pairs) and continue to train for 40K iterations. In stage 2, we train the 1.5B model using 9M multimodal instructional and 16M high-quality generation data for a total of around 35K iterations. α in Eq. 4 is set as 1.0. The stage 2 training process takes around 15 hours. For models with mixed-modality and video generation capabilities, we progressively add video-text and interleaved data in stage 1. For video data, we randomly sample a 2s 480p or 432×432 clip with 17 frames from each video with an interval of 3 frames. The context length at this time is set as 7006. In stage 2, high-quality video-text and interleaved data are added to further improve video and mixed-modality generation capabilities.

In the training of our model based on the 7B LLM variant, we resume the flow head pre-trained based on the 1.5B model and additionally train the newly initialized spatial (-temporal) fusion, projector, and MLP transformations for 3K iterations with 2K warm-up steps to align the hidden size and then further train spatial (-temporal) fusion, the projector, MLP transformations, and the flow head together. Following that, we conduct the training stages 1 and 2 in the same manner as those of the 1.5B model. The whole training process of our 7B model takes approximately 2 and a half days using 128 H100 GPUs. We do not include interleaved and video data in the training stages of the larger model due to the huge computational cost and training duration.

Table 4: Evaluation on the GenEval [37] benchmark. Gen. denotes "generation". # Params. indicates the number of parameters of base LLM. # Data. indicates the number of image-text pairs used for visual generation during training. * indicates the method uses two distinct models or sets of parameters for multimodal understanding and generation, respectively. Obj.: Object. Attri.: Attribute. Our results are obtained based on the rewritten dense prompts.

Type	Method	# Params.	# Data	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	**Overall↑**
Gen. Only	SD3 (d=24) [33]		-	0.98	0.74	0.63	0.67	0.34	0.36	0.62
	SD3-Medium [33]	-	-	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Unifying via Tailored Models	SEED-X [36]	17B	158M	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	TokenFlow-XL* [77]	14B	60M	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	ILLUME [92]	7B	15M	0.99	0.86	0.45	0.71	0.39	0.28	0.61
Native Unified	Show-o [104]	1.3B	2.0B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
	Emu3 [94]	8B	-	-	-	-	-	-	-	0.66
	MUSE-VL [105]	7B	24M							0.57
	Transfusion [117]	7B	3.5B	-	-	-	-	-	-	0.63
	D-DiT [57]	2B	40M	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	Janus-Pro [23]	7B	144M	0.99	0.89	0.59	0.90	0.79	0.66	0.80
	Mogao [59]	7B	-	1.00	0.97	0.83	0.93	0.84	0.80	0.89
	Show-o2 (Ours)	1.5B	66M	0.99	0.86	0.55	0.86	0.46	0.63	0.73
	Show-o2 (Ours)	7B	66M	1.00	0.87	0.58	0.92	0.52	0.62	0.76

Table 5: Evaluation on the DPG-Bench [40] benchmark. Gen. denotes "generation". # Params. indicates the number of parameters of base LLM. # Data. indicates the number of image-text pairs used for visual generation during training.

Type	Method	# Params.	# Data	Global	Entity	Attribute	Relation	Other	**Overall↑**
Gen. Only	Hunyuan-DiT [58]	1.5B	-	84.59	80.59	88.01	74.36	86.41	78.87
	Playground v2.5 [51]	-	-	83.06	82.59	81.20	84.08	83.50	75.47
	PixArt-Σ [15]	-	-	86.89	82.89	88.94	86.59	87.68	80.54
	DALL-E 3 [9]	-	-	90.97	89.61	88.39	90.58	89.83	83.50
	SD3-Medium [33]	2B	-	87.90	91.01	88.83	80.70	88.68	84.08
Native Unified	Emu3-DPO [94]	8B	-	-	-		-	-	81.60
	Janus-Pro [23]	7B	144M	86.90	88.90	89.40	89.32	89.48	84.19
	Mogao [59]	7B	-	82.37	90.03	88.26	93.18	85.40	84.33
	Show-o2 (Ours)	1.5B	66M	87.53	90.38	91.34	90.30	91.21	85.02
	Show-o2 (Ours)	7B	66M	89.00	91.78	89.96	91.81	91.64	**86.14**

4.2 Multimodal Understanding

Quantitative Results. Table 3 highlights the performance of our models on multimodal understanding benchmarks, evaluated across metrics such as MME [35], GQA [44], SEED-Bench [49], MM-Bench [66], MMU [110], MMStar [19], and AI2D [46]. As shown in the table, both the 1.5B and 7B variants of our model consistently outperform state-of-the-art models across many metrics. For models with similar parameter sizes (1.5B), our model achieves the best scores on MME-p and MMU-val benchmarks while delivering competitive performance on GQA and SEED-Bench metrics. When compared to larger models with approximately 7B parameters, our models surpass state-of-the-art models such as Janus-Pro and even the significantly larger TokenFlow-XL model (14B parameters) in metrics including MME-p, GQA, MMMU-val, MMStar, and AI2D, while maintaining competitive performance on SEED-Bench and MM-Bench. These results underscore the robust perception capabilities of our unified visual representations, demonstrating their effectiveness in multimodal understanding tasks and the promising potentials in this domain.

Qualitative Results. Fig. 3 showcases the multimodal understanding capabilities of our model. As demonstrated, the model excels at answering general-purpose questions about an image. Specifically, it can provide detailed descriptions of an image, count objects, and recognize text within the image. Besides, the model can leverage its world knowledge to offer step-by-step instructions for preparing daily drinks like an avocado milkshake and supports bilingual question-answering, highlighting its versatility and practical utility. Further, our model supports multimodal understanding in both English and Chinese, enabling bilingual capabilities.

4.3 Visual Generation

Image Generation. We compare our model with the state-of-the-art approaches on GenEval [37] and DPG-Bench [40] benchmarks in Tables 4 and 5. One can observe that our model surpasses most approaches, including TokenFlow-XL, Show-o, Emu3, and Transfusion, on the GenEval benchmark. Compared to Janus-Pro, which was trained on a significantly larger dataset of 144M image-text pairs,

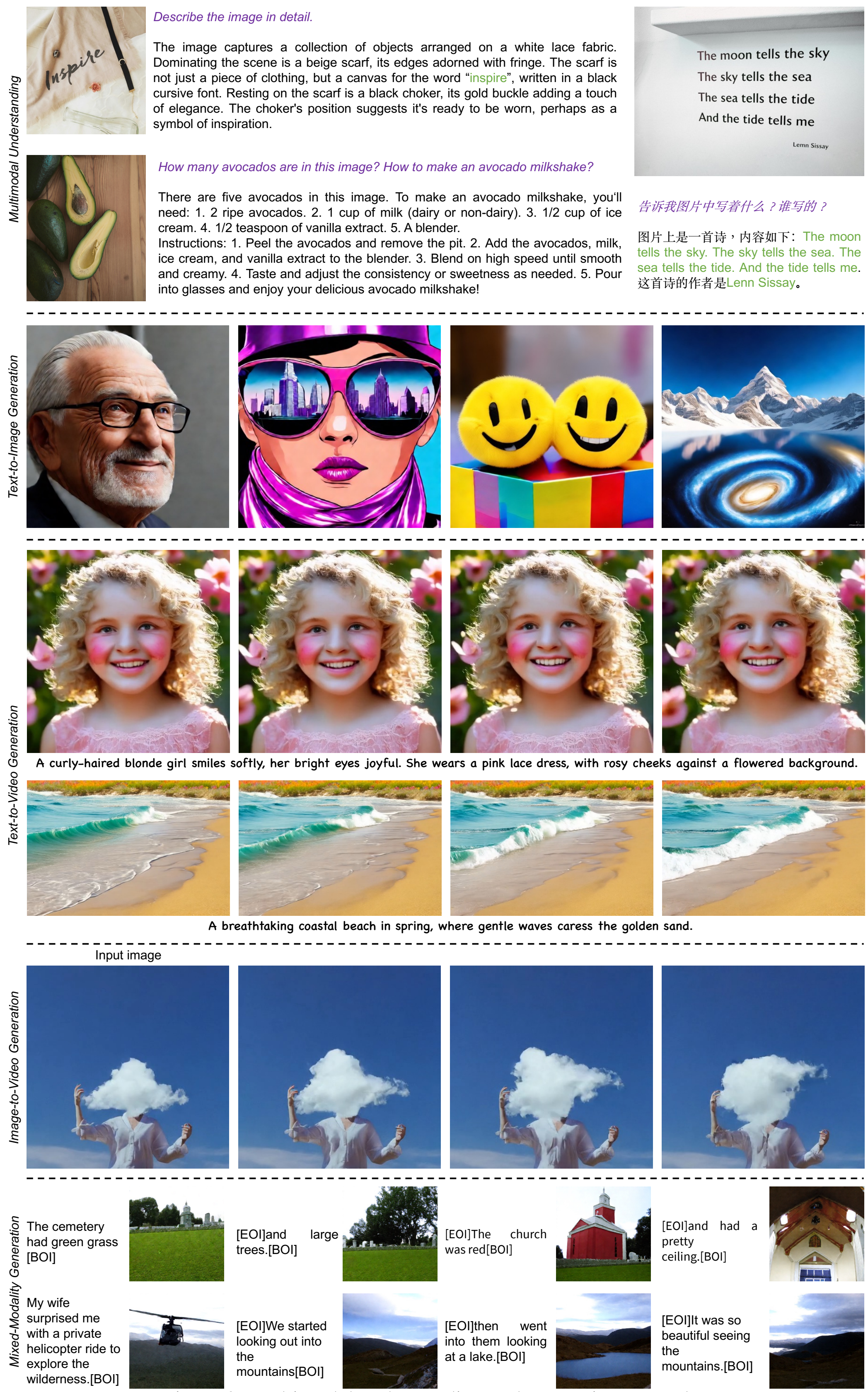

Figure 2: Multimodal understanding and generation examples.

Models	# Params.	Total	QS	SS	SC	BC	TF	MS	DD	AQ	IQ	OC	MO	HA	C	SR	S	AS	TS	OC’
ModelScope [93]	1.7B	75.75	78.05	66.54	89.87	95.29	98.28	95.79	66.39	52.06	58.57	82.25	38.98	92.40	81.72	33.68	39.26	23.39	25.37	25.67
LaVie [95]	3B	77.08	78.78	70.31	91.41	97.47	98.30	96.38	49.72	54.94	61.90	91.82	33.32	96.80	86.39	34.09	52.69	23.56	25.93	26.41
OpenSoraPlan V1.3 [60]	-	77.23	80.14	65.62	97.79	97.24	99.20	99.05	30.28	60.42	56.21	85.56	43.58	86.80	79.30	51.61	36.73	20.03	22.47	24.47
Show-1 [113]	6B	78.93	80.42	72.98	95.53	98.02	99.12	98.24	44.44	57.35	58.66	93.07	45.47	95.60	86.35	53.50	47.03	23.06	25.28	27.46
AnimateDiff-V2 [39]	-	80.27	82.90	69.75	95.30	97.68	98.75	97.76	40.83	67.16	70.10	90.90	36.88	92.60	87.47	34.60	50.19	22.42	26.03	27.04
Gen-2 [1]	-	80.58	82.47	73.03	97.61	97.61	99.56	99.58	18.89	66.96	67.42	90.92	55.47	89.20	89.49	66.91	48.91	19.34	24.12	26.17
Pika-1.0 [2]	-	80.69	82.92	71.77	96.94	97.36	99.74	99.50	47.50	62.04	61.87	88.72	43.08	86.20	90.57	61.03	49.83	22.26	24.22	25.94
VideoCrafter-2.0 [14]	-	80.44	82.20	73.42	96.85	98.22	98.41	97.73	42.50	63.13	67.22	92.55	40.66	95.00	92.92	35.86	55.29	25.13	25.84	28.23
CogVideoX [109]	5B	81.61	82.75	77.04	96.23	96.52	98.66	96.92	70.97	61.98	62.90	85.23	62.11	99.40	82.81	66.35	53.20	24.91	25.38	27.59
Kling [4]	-	81.85	83.39	75.68	98.33	97.60	99.30	99.40	46.94	61.21	65.62	87.24	68.05	93.40	89.90	73.03	50.86	19.62	24.17	26.42
Step-Video-T2V [69]	30B	81.83	84.46	71.28	98.05	97.67	99.40	99.08	53.06	61.23	70.63	80.56	50.55	94.00	88.25	71.47	24.38	23.17	26.01	27.12
Gen-3 [3]	-	82.32	84.11	75.17	97.10	96.62	98.61	99.23	60.14	63.34	66.82	87.81	53.64	96.40	80.90	65.09	54.57	24.31	24.71	26.69
Emu3 [94]	8B	80.96	-	-	95.32	97.69	-	98.93	79.27	59.64	-	86.17	44.64	77.71	-	68.73	37.11	20.92	-	-
VILA-U [101]	7B	74.01	76.26	65.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Show-o2	2B	81.34	82.10	78.31	97.28	96.78	97.68	98.25	40.83	65.15	67.06	94.81	76.01	95.20	80.89	62.61	57.67	23.29	25.27	27.00

Table 6: Comparison with text-to-video models on the VBench [43] benchmark. # Params. indicates the number of total parameters for video generation. QS: Quality Score, SS: Semantic Score, SC: Subject Consistency, BC: Background Consistency, TF: Temporal Flickering, MS: Motion Smoothness, DD: Dynamic Degree, AQ: Aesthetic Quality, IQ: Imaging Quality, OC: Object Class, MO: Multiple Objects, HA: Human Action, C: Color, SR: Spatial Relationship, S: Scene, AS: Appearance style, TS: Temporal Style, OC’: Overall Consistency.

Models	I2V Subject	I2V Background	Camera Motion	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
DynamiCrafter-1024 [106]	96.71	96.05	35.44	95.69	97.38	97.63	97.38	47.40	66.46	69.34
SEINE-512x320 [22]	94.85	94.02	23.36	94.20	97.26	96.72	96.68	34.31	58.42	70.97
I2VGen-XL [115]	96.74	95.44	13.32	96.36	97.93	98.48	98.31	24.96	65.33	69.85
Animate-Anything [27]	98.54	96.88	12.56	98.90	98.19	98.14	98.61	2.68	67.12	72.09
ConsistI2V [79]	94.69	94.57	33.60	95.27	98.28	97.56	97.38	18.62	59.00	66.92
VideoCrafter-I2V [13]	90.97	90.51	33.58	97.86	98.79	98.19	98.00	22.60	60.78	71.68
SVD-XT-1.1 [10]	97.51	97.62	-	95.42	96.77	99.17	98.12	43.17	60.23	70.23
MarDini [64]	98.78	96.46	-	-	-	-	-	-	-	-
Show-o2	96.94	98.83	28.41	93.83	97.45	-	97.76	25.85	61.92	69.87

Table 7: Comparison with image-to-video models on the VBench [43] benchmark.

our model achieves promising results with only 66M image-text pairs. On DPG-Bench evaluation, our model has demonstrated the best overall score compared to generation-only models such as SD3-Medium and unified models, including Emu3-DPO and Janus-Pro. We also show qualitative results in Fig. 3 to illustrate that our model can generate high-quality and realistic images.

Video Generation. We compare our model with the text-to-video and image-to-video generation models in Tables 6 and 7. One can observe that with only 2B parameters, our model outperforms models such as Show-1, Emu3, and VILA-U with more than 6B parameters. Besides, our model has demonstrated competitive performance compared to CogVideoX and Step-Video-T2V. We also provide qualitative results of the text-to-video and image-to-video generation capability of our model in the middle of Fig. 3. One can observe that, given text prompts or an input image, our model can generate consistent video frames with reasonable motions, such as the smiling girl, lapping waves, and floating clouds.

4.4 Mixed-Modality Generation

We demonstrate mixed-modality generation capabilities of our model using downstream task visual storytelling dataset [42] in Fig. 3. During fine-tuning, given an interleaved image-text sequence, we apply noise to all images in the sequence with a probability of 0.3. Otherwise, we randomly retain a number of the earlier images in the sequence and only apply noise to the later ones. Benefiting from the general interleaved sequence format mentioned in 3.1, our model can predict the [BOI] once it begins to generate an image. Upon detecting the [BOI] token, noises will be appended to the sequence to gradually generate an image. The generated text tokens and images will be served as context to continue generating the following output. Fig. 3 includes two examples demonstrating our model’s ability to interleavely generate coherent text and images, vividly narrating a story.

4.5 Ablation Studies

We show the pilot study results in Table 8, which validated the effect of spatial (-temporal) fusion on multimodal understanding and generation performance. For efficiency, we adopt LLaMA-3.2-1B as the base language model and use only around 1M multimodal understanding data and the ImageNet-1K generation data [29]. Under the same training settings, there are improvements in terms of both multimodal understanding and generation metrics, including MME-p, GQA, and FID-5K. This validates that the involved semantic and low-level features in the fusion mechanism would potentially help both the multimodal generation and understanding capabilities to some extent.

Table 8: Effect of spatial (-temporal) fusion.

	MME−p ↑	GQA ↑	POPE ↑	FID-5K ↓
w/o Fusion	1164.7	56.2	82.6	21.8
w Fusion	**1187.8**	**57.6**	82.6	**20.5**

We perform ablation studies to examine the effect of classifier-free guidance (CFG) and inference steps on the generative performance using the 1.5B model. One can observe that increasing the CFG guidance scale and inference steps (in a range) would potentially improve the GenEval and DPG-Bench scores. However, the improvements of the GenEval score are not significant when the CFG guidance is set as larger than 5.0.

Table 9: Effect of CFG guidance and inference steps.

CFG guidance	Inference steps	GenEval	DPG-Bench
2.5	50	0.65	81.6
5.0	50	0.71	83.9
7.5	50	0.71	84.8
10	50	0.71	**85.0**
7.5	25	0.71	84.6
7.5	100	**0.73**	84.7

Table 10 provides the effect of training stages on the generation performance on the GenEval and DPG-Bench benchmarks. One can observe that stage-2 training consistently and significantly improves both metrics, which validates the importance of the second stage.

Table 10: Effect of training stages.

Stage-1	Stage-2	GenEval	DPG-Bench
✓		0.63	83.28
✓	✓	**0.73**	**84.70**

5 Limitations and Broader Impacts

We found that our model is not good at rendering text on the image. We investigated our generation datasets and observed that the proportion of images with rendered texts is relatively small, which potentially leads to bad text rendering. In addition, the generated images will lack details of the small objects because of the limited image resolution.

Our models possess the ability to generate text and images, which may carry the risk of unintended misuse, such as creating fake information or profiles. Additionally, our large-scale dataset includes content featuring celebrities and copyrighted materials, which could potentially result in intellectual property infringement.

6 Conclusion

This paper proposed improved native unified multimodal models scalable for multimodal understanding and generation, image and video modalities, by integrating 3D causal VAE, autoregressive modeling, and flow matching. A dual-path of spatial (-temporal) fusion mechanism guided the construction of unified visual representations with both high- and low-level features. A two-stage training recipe enables effective learning of unified capabilties, resulting in a versatile model capable of handling diverse tasks, including multimodal understanding and image/video generation. Extensive experiments demonstrate the model's state-of-the-art performance across various benchmarks.

Acknowledgments and Disclosure of Funding

We appreciate Haozhe Liu for his consistent input and discussions regarding this project.

Table 11: Comparative analysis of selected unified multimodal models based on the utilization of visual representations and type of unified modeling for multimodal understanding and generation. In this context, **native und. & gen.** refers to the direct decoding of output sequences into texts and images, as opposed to serving as conditions for decoding using external pre-trained decoders like Stable Diffusion. Please refer to the complete table in the appendix. * indicates the method uses two distinct models for multimodal understanding and generation, respectively.

Methods	Und. & Gen. Representation			Type of Unified Modeling		
	Unified	Decoupled	Support Video	Native Und. & Gen.	Assembling Tailored Models	Paradigm
Chameleon [86]	✓		✗	✓		AR
Show-o [104]	✓		✗	✓		AR + Diff.
Transfusion [117]	✓		✗	✓		AR + Diff.
VILA-U [101]	✓		✓	✓		AR
Emu3 [94]	✓		✓	✓		AR
MonoFormer [116]	✓		✗	✓		AR + Diff.
Dual-Diffusion [57]	✓		✗	✓		Diff.
SynerGen-VL [52]	✓		✗	✓		AR
MMAR [108]	✓		✗	✓		AR + MAR
MUSE-VL [105]	✓		✗	✓		AR
Orthus [48]	✓		✗	✓		AR + Diff.
Liquid [97]	✓		✗	✓		AR
UGen [84]	✓		✗	✓		AR
UniToken [45]	✓		✗	✓		AR
Harmon [100]	✓		✗	✓		AR+MAR
DualToken [82]	✓		✗	✓		AR
UniTok [68]	✓		✗	✓		AR
VARGPT [118]	✓		✗	✓		AR
Selftok [91]	✓		✗	✓		AR
Show-o2 (Ours)	✓		✓	✓		AR + Diff.
Janus-Series [23, 24, 70]		✓	✗	✓		AR (+Diff.)
UnidFluid [34]		✓	✗	✓		AR + MAR
OmniMamba [119]		✓	✗	✓		AR
Mogao [59]		✓	✗	✓		AR + Diff.
Bagel [28]		✓	✓	✓		AR + Diff.
NExT-GPT [99]		✓	✓		✓	AR + Diff.
SEED-X [36]		✓	✗		✓	AR + Diff.
MIO [96]		✓	✓		✓	AR + Diff.
MetaMorph [88]		✓	✗		✓	AR + Diff.
ILLUME [92]		✓	✗		✓	AR + Diff.
ILLUME+ [41]		✓	✗		✓	AR + Diff.
MetaQueries [72]		✓	✗		✓	AR + Diff.
Nexus-Gen [114]		✓	✗		✓	AR + Diff.
Ming-Lite-Uni [38]		✓	✗		✓	AR + Diff.
BLIP3-o [16]		✓	✗		✓	AR + Diff.
TokenFlow* [77]	✓		✗		✓	AR
LlamaFusion [81]	✓		✗	✓		AR + Diff.
SemHiTok* [25]	✓		✗		✓	AR

A Technical Appendices and Supplementary Material

A.1 More Qualitative Results

Figure 3: Multimodal understanding and generation examples.

A.2 Text Prompts

We provide the text prompts used in Fig. 3 below:

"An elderly man, seemingly in his 70s or 80s, captured in stunning high-definition detail. His face is adorned with a neatly groomed white beard and mustache, each strand visible and adding a sense of wisdom and experience to his appearance. He wears sleek black glasses, the frames polished and reflecting light subtly, enhancing his intellectual aura. His skin shows the fine lines and wrinkles of age, each crease telling a story of time and life lived. His gaze is directed upwards and slightly to the left, with his deep-set eyes conveying a sense of contemplation or curiosity, as if he is pondering something profound or observing an unseen detail beyond the frame. The background is a smooth, neutral gray wall, its texture faintly visible, ensuring the focus remains entirely on the man. The lighting is soft yet precise, highlighting every feature of his face, from the texture of his skin to the glint in his glasses, creating a portrait rich in depth and character. "

"A digital illustration featuring a close-up of a person's face wearing large sunglasses. The sunglasses reflect an urban landscape with skyscrapers, creating a striking visual effect. The person is also wearing a shiny pink-purple scarf or hat, adding richness and vibrant color to the image. The colors are bright and saturated, evoking a futuristic impression. The overall style combines elements of fashion, dark fantasy, conceptual art, and vibrant architecture. "

"Two bright yellow plush smiley faces sitting side by side in a colorful, rainbow-colored box. The plush toys have cheerful expressions, with one smiling widely and the other grinning with a slightly open mouth. The box is simple yet vibrant, featuring a rainbow-colored design without any additional patterns or decorations. The background of the scene is blurred, drawing attention to the playful and happy vibe of the plush toys and the colorful box. The overall composition exudes positivity, fun, and a sense of vibrant energy. "

"A photograph of the Mont Blanc mountain range, showcasing an incredible view from Aiguille du Midi. The snow-covered peaks of the Swiss Alps dominate the frame, bathed in the crisp light of a vibrant blue sky, while a subtle mirrored reflection of the scene extends into the vastness of space. Below the mountains, a cosmic swirl of blue and gold galaxies is visible, hinting at a dissolving form of energy and a new understanding, creating a breathtaking vista of the Alps in their purest form. Soft, ethereal lighting highlights the peaks and enhances the overall sense of awe and wonder. "

References

[1] Gen-2. Accessed September 25, 2023 [Online] https://research.runwayml.com/gen2, 2023.

[2] Pika 1.0. Accessed December 28, 2023 [Online] https://www.pika.art/, 2023.

[3] Gen-3. Accessed June 17, 2024 [Online] https://runwayml.com/research/introducing-gen-3-alpha, 2024.

[4] Kling. Accessed June 6, 2024 [Online] https://klingai.kuaishou.com/, 2024.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[8] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

[9] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

[10] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[11] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[12] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021.

[13] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[14] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[15] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.

[16] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[17] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*. OpenReview.net, 2024.

[18] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[19] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[20] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703, 2020.

[21] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[22] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023.

[23] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[24] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[25] Zisheng Chen, Chunwei Wang, Xiuwei Chen, Hang Xu, Jianhua Han, and Xiaodan Liang. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. *arXiv preprint arXiv:2503.06764*, 2025.

[26] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[27] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Fine-grained open domain image animation with motion guidance. *arXiv preprint arXiv:2311.12886*, 2023.

[28] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[30] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

[31] Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. *arXiv preprint arXiv:2502.06788*, 2025.

[32] 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. DreamLLM: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.

[33] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[34] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

[35] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

[36] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[37] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

[38] Biao Gong, Cheng Zou, Dandan Zheng, Hu Yu, Jingdong Chen, Jianxin Sun, Junbo Zhao, Jun Zhou, Kaixiang Ji, Lixiang Ru, et al. Ming-lite-uni: Advancements in unified architecture for natural multimodal interaction. *arXiv preprint arXiv:2505.02471*, 2025.

[39] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

[40] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

[41] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, and Hang Xu. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.

[42] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016.

[43] 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. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[44] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019.

[45] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. *arXiv preprint arXiv:2504.04423*, 2025.

[46] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[47] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[48] Siqi Kou, Jiachun Jin, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024.

[49] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[50] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[51] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *ArXiv*, abs/2402.17245, 2024.

[52] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024.

[53] Haopeng Li, Jinyue Yang, Guoqi Li, and Huan Wang. Autoregressive image generation with randomized parallel decoding. *arXiv preprint arXiv:2503.10568*, 2025.

[54] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. In *The Thirteenth International Conference on Learning Representations*, 2025.

[55] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

[56] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *2407.08303*, 2024.

[57] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024.

[58] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[59] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

[60] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[61] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[62] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[63] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024.

[64] Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C. Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, Jui-Chieh Wu, Sen He, Tao Xiang, Jürgen Schmidhuber, and Juan-Manuel Pérez-Rúa. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv:2410.20280*, 2024.

[65] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[66] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[67] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

[68] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

[69] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025.

[70] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024.

[71] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[72] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[73] Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024.

[74] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[75] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[76] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Wenhai Wang Victor Perez, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework, 2025.

[77] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

[78] 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. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[79] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[81] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

[82] Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*, 2025.

[83] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[84] Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025.

[85] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *NeurIPS*, 36, 2024.

[86] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[87] 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. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

[88] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

[89] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[90] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[91] Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, Mingze Zhou, Wang Lin, Kaihang Pan, Saining Zhang, Liyu Jia, Wentao Hu, Wei Zhao, and Hanwang Zhang. Discrete visual tokens of autoregression, by diffusion, and for reasoning. 2025.

[92] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. ILLUME: illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024.

[93] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[94] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[95] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2024.

[96] 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. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv: 2409.17692*, 2024.

[97] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024.

[98] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023.

[99] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[100] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation, 2025.

[101] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[102] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer, 2025.

[103] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, pages 7452–7461, 2023.

[104] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.

[105] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. MUSE-VL: modeling unified VLM through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024.

[106] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[107] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[108] Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. MMAR: towards lossless multi-modal auto-regressive probabilistic modeling. *arXiv preprint arXiv:2410.10798*, 2024.

[109] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[110] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pages 9556–9567. IEEE, 2024.

[111] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

[112] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.

[113] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

[114] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, Yuze Zhao, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025.

[115] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[116] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024.

[117] Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025.

[118] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model, 2025.

[119] Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models, 2025.