OmniMamba: Efficient and Unified Multimodal Understanding and Generation via State Space Models

Jialv Zou[1], Bencheng Liao[2,1], Qian Zhang[3], Wenyu Liu[1], and Xinggang Wang[1,✉]

[1] School of Electronic Information and Communications, Huazhong University of Science & Technology, Wuhan, China
[2] Institute of Artificial Intelligence, Huazhong University of Science & Technology, Wuhan, China
[3] Horizon Robotics, Beijing, China

Abstract. Recent advancements in unified multimodal understanding and visual generation (or multimodal generation) models have been hindered by their quadratic computational complexity and dependence on large-scale training data. We present OmniMamba, the first linear-architecture-based multimodal generation model that generates both text and images through a unified next-token prediction paradigm. The model fully leverages Mamba-2's high computational and memory efficiency, extending its capabilities from text generation to multimodal generation. To address the data inefficiency of existing unified models, we propose two key innovations: (1) decoupled vocabularies to guide modality-specific generation, and (2) task-specific LoRA for parameter-efficient adaptation. Furthermore, we introduce a decoupled two-stage training strategy to mitigate data imbalance between two tasks. Equipped with these techniques, OmniMamba achieves competitive performance with JanusFlow while surpassing Show-o across benchmarks, despite being trained on merely 2M image-text pairs, which is 1,000 times fewer than Show-o. Notably, OmniMamba stands out with outstanding inference efficiency, achieving up to a 119.2× speedup and 63% GPU memory reduction for long-sequence generation compared to Transformer-based counterparts. Code and models are released at https://github.com/hustvl/OmniMamba.

Keywords: Unified Model · Mamba · State Space Models

1 Introduction

In recent years, Large Language Models (LLMs) [3, 6, 16, 61, 62] have achieved remarkable advancements, igniting significant research interest in extending their fundamental capabilities to the visual domain. Consequently, researchers have developed a series of Multimodal Large Language Models (MLLMs) for tasks such as multimodal understanding [44, 45, 77, 79] and visual generation [32, 57].

✉ Corresponding author: xgwang@hust.edu.cn

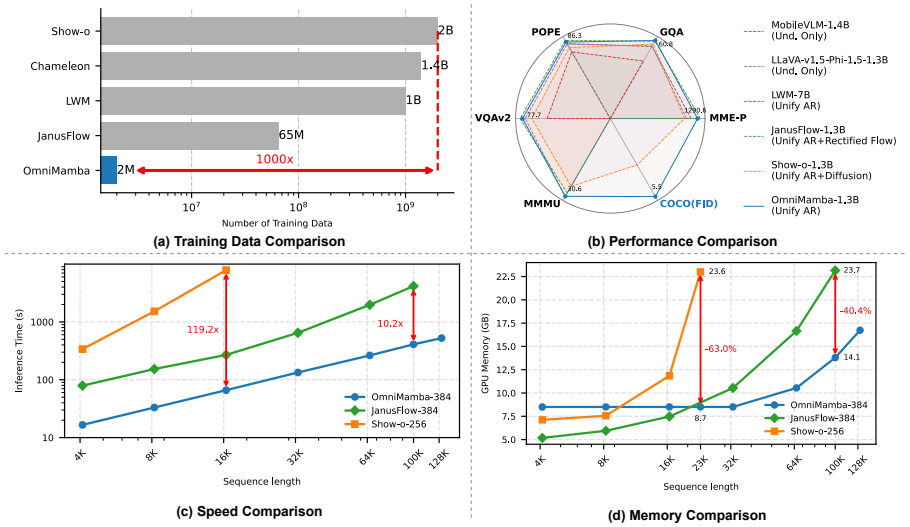

Fig. 1: Comprehensive comparison between OmniMamba and other unified understanding and generation models. (a) Our OmniMamba is trained on only 2M image-text pairs, which is 1000 times less than Show-o. **(b)** With such limited data for training, our OmniMamba significantly outperforms Show-o across a wide range of benchmarks and achieves competitive performance with JanusFlow. Black metrics are for the multimodal understanding benchmark, while the blue metric is for the visual generation task. **(c)-(d)** We compare the speed and memory of OmniMamba with other unified models on the same single NVIDIA 4090 GPU. OmniMamba demonstrates up to a 119.2× speedup and 63% GPU memory reduction for long-sequence generation.

Recent studies have emerged that seek to integrate multimodal understanding with visual generation, aiming to develop unified systems capable of handling both tasks simultaneously. Such designs hold the potential to foster mutual enhancement between generation and understanding, offering a promising pathway toward truly unifying all modalities. Numerous studies have sought to preserve the text generation paradigm of LLMs while exploring the impact [47, 66, 68, 69] of integrating diverse visual generation paradigms, such as diffusion models [25], flow-based generative models [17, 41], and vector-quantized autoregressive models [58].

Unfortunately, the significant domain gap between image and text presents a critical challenge for unified multimodal generative models: preserving generation capabilities without degrading understanding performance requires an extensive volume of image-text pairs for training, as illustrated in Fig. 1. This not only leads to poor training efficiency but also creates a substantial barrier to the broader development of such models, as only a small fraction of researchers possess the resources to undertake such computationally demanding studies. Moreover, most existing unified multimodal generative models rely on Transformer-based LLMs [63]. However, their quadratic computational complex-

ity results in slow inference speeds, rendering them less practical for real-time applications.

The challenges faced by existing unified multimodal generative models naturally lead us to ponder: **can a model be developed that achieves both training efficiency and inference efficiency?**

To address this, we introduce OmniMamba, a novel unified multimodal generative model that requires only 2M image-text pairs for training. Built on the Mamba-2-1.3B [11] model as the foundational LLM with a unified next token prediction paradigm to generate all modalities, OmniMamba leverages the linear computational complexity of state space models (SSMs) to achieve significantly faster inference speeds. Furthermore, to empower the Mamba-2 LLM—whose foundational capabilities are relatively weaker compared to the extensively studied Transformer models—to efficiently learn mixed-modality generation with limited training data, we propose novel model architectures and training strategies.

To enhance the model's capability in handling diverse tasks, we incorporate task-specific LoRA [26]. Specifically, within each Mamba-2 layer's input linear projection, we introduce distinct LoRA modules for multimodal understanding and visual generation. During task execution, the features are modulated by both the linear projection and the corresponding task-specific LoRA, while the irrelevant LoRA components are deactivated. Furthermore, we propose the decoupled vocabularies to guide the model in generating the appropriate modality, which requires more data for the model to learn. On the data front, we further propose a novel two-stage decoupled training strategy to address the data imbalance between the two tasks, significantly improving training efficiency.

Trained on only 2M image-text pairs, our proposed OmniMamba outperforms Show-o [69] on multiple multimodal understanding benchmarks and also matches the performance of JanusFlow [47], which was introduced by DeepSeek AI. Moreover, it achieves the best visual generation performance on the MS-COCO dataset [40]. Notably, OmniMamba demonstrates a 119.2× speedup at a sequence length of 16k and a 63% GPU memory reduction at a sequence length of 23k, compared to Show-o. Furthermore, at a sequence length of 100k, it achieves a 10.2× speedup and 40.4% memory savings compared to JanusFlow. Our main contributions can be summarized as follows:

- We introduce OmniMamba, the first Mamba-based unified multimodal understanding and visual generation model to the best of our knowledge. By novelly adopting decoupled vocabularies and task-specific LoRA, OmniMamba achieves effective training and inference.
- We propose a novel decoupled two-stage training strategy to address the issue of data imbalance between tasks. With this strategy and our model design, OmniMamba achieves competitive performance using only 2M image-text pairs for training-up to 1,000 times fewer than previous SOTA models.
- Comprehensive experimental results show that OmniMamba achieves competitive or even superior performance across a wide range of vision-language benchmarks and MS-COCO generation benchmark, with significantly im-

proved inference efficiency, achieving up to a 119.2× speedup and 63% GPU memory reduction for long-sequence generation on NVIDIA 4090 GPU.

2 Related Work

Multimodal Understanding The remarkable advancements in LLMs have catalyzed the development of Large Vision-Language Models (LVLMs). Some representative works, such as the LLaVA series [45,81], BLIP series [36,37], and MiniGPT-4 [79], have demonstrated strong multimodal understanding capabilities. These models align the features obtained from pretrained vision encoders with the feature space of LLMs through feature projectors, enabling pretrained LLMs to transfer their understanding and reasoning abilities to multimodal scenarios.

Visual Generation In recent years, diffusion models [25] have made remarkable progress, leading to models [13,50,56] with strong visual generation capabilities. Building on these advancements, flow-based generative models [17,41] have achieved superior results with fewer sampling steps. Additionally, some works [58,73] have successfully integrated autoregressive models into this domain, achieving notable performance.

Unified Understanding and Generation The remarkable advancements of LLMs in the fields of multimodal understanding and visual generation have naturally sparked researchers' interest in training a single LLM for both tasks. Early works [14,19–21] integrated pretrained diffusion modules as tools into LLMs, essentially forming a combination of two expert systems rather than utilizing a single LLM to perform both tasks. This approach results in a more complex model architecture and often leads to suboptimal outcomes. Show-o [69] integrates next-token prediction with discrete diffusion, enabling adaptive handling of mixed-modality inputs and outputs. Meanwhile, JanusFlow [47] merges autoregressive models with rectified flow, a cutting-edge technique in visual generation. In contrast, Emu3 [66] asserts that next-token prediction holds the greatest potential for achieving multi-modal generation, relying exclusively on this paradigm to manage both text and image generation tasks. Although these methods achieve outstanding performance, they are all based on Transformers, whose quadratic computational complexity presents a significant drawback, particularly when handling long-sequence generation tasks and high-resolution image generation. To address this challenge, we propose OmniMamba, which employs a unified next token prediction paradigm to generate both text and image modalities, aiming to extend the linear computational complexity of the linear models to the field of multimodal generation.

Linear Model In recent years, a series of linear-complexity models have emerged as strong competitors to Transformers. Mamba [22], a selective state space model, has garnered widespread attention for its competitive performance and

faster inference speeds compared to Transformers. Building on this, Mamba-2 [11], an enhanced version of Mamba, achieves performance on par with Transformers while being 2-8 times faster. Similarly, GLA [70] leverages gated linear attention to achieve linear complexity, maintaining competitive performance with Transformers.

The success of linear-complexity models in the field of natural language processing (NLP) has inspired its application in the visual domain, where it has been extensively studied in traditional image tasks [46, 80], multimodal understanding [28, 39, 51, 76], and visual generation [27, 35]. In this paper, we aim to design a unified multimodal understanding and visual generation model based on Mamba-2, which maintains competitive performance with Transformer-based unified models while offering significantly faster inference speeds.

3 Method

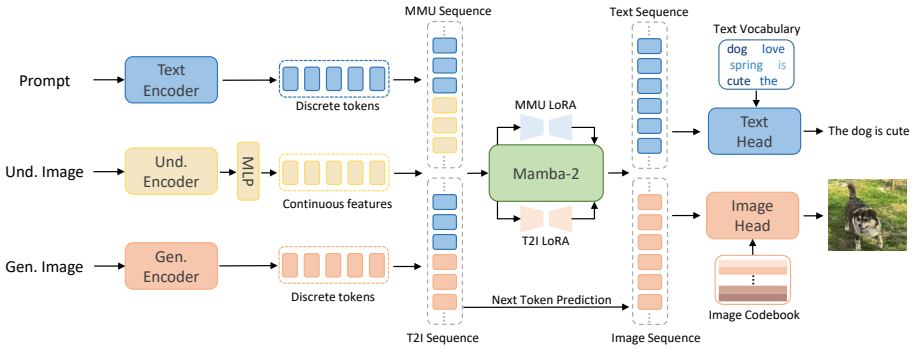

Fig. 2: Architecture of the proposed OmniMamba "MMU" refers to multimodal understanding, while "T2I" refers to text-to-image generation. OmniMamba employs a next-token prediction paradigm for both multimodal understanding and visual generation tasks. To address the distinct requirements of each task—semantic information extraction for multimodal understanding and high-fidelity image compression for visual generation, we utilize separate encoders and heads. Furthermore, we purpose decoupled vocabularies to guide modality-specific generation and task-specific LoRA for parameter-efficient adaptation.

3.1 Overall Architecture

Our ultimate goal is to design a unified multimodal understanding and visual generation model that achieves both training and inference efficiency using only 2M image-text pairs for training. We believe the key to realizing this goal can be summarized in one word: **decoupling**. To this end, we propose OmniMamba, the architecture of which is illustrated in Fig. 2.

Success necessitates standing on the shoulders of giants. We observe Emu3 [66], an autoregressive-based model which employs vast amounts of data and 8 billion model parameters. Despite these advantages, its final performance remains

suboptimal, falling short of JanusFlow [47], a hybrid generative paradigm-based model with significantly less data and fewer parameters. We argue that this discrepancy stems not from the inherent superiority of the hybrid generative paradigm but from Emu3's tight coupling design, it uses the same vocabulary and encoder for all tasks and modalities. While this design aligns with the original intention of a unified model, it may lead to inefficient data utilization. In the following, we will introduce our model by focusing on the concept of decoupling.

3.2 Decoupling Encoders for the Two Tasks

Previous works have explored using a single vision encoder for both tasks. For example, Show-o [69] employs MAGVIT-v2 [72] to encode images into discrete tokens for both understanding and generation tasks. TransFusion [78] utilizes a shared U-Net or a linear encoder to map images into a continuous latent space for both tasks. Emu3 trains its vision encoder based on SBER-MoVQGAN, enabling the encoding of video clips or images into discrete tokens.

However, JanusFlow [47] has shown that such a unified encoder design is suboptimal. We believe this is primarily because multimodal understanding requires rich semantic representations for complex reasoning, whereas visual generation focuses on precisely encoding the spatial structure and texture of images. The inherent conflict between these two objectives suggests that a unified encoder design may not be the optimal choice. Therefore, OmniMamba adopts a decoupled vision encoder design.

Following prismatic VLMs [31], we fuse DINOv2 [49] and SigLIP [75] as an encoder to extract continuous features for multimodal understanding. The key idea is that integrating visual representations from DINOv2, which capture low-level spatial properties, with the semantic features provided by SigLIP leads to further performance improvements. For visual generation, we use an image tokenizer trained with LlamaGen [58] to encode images into discrete representations. This tokenizer was pretrained on ImageNet [12] and further fine-tuned on a combination of 50M LAION-COCO [34] and 10M internal high aesthetic quality data.

3.3 Decoupling Vocabularies for the Two Tasks

Unlike Emu3 and Show-o, which use a large unified vocabulary to represent both text and image modalities, to disentangle modality-specific semantics, we employ two separate vocabularies for each modality. This design explicitly separates the two modalities, providing additional modality-level prior knowledge. As a result, the model does not need to learn whether the output should be text or image, instead, it ensures the correct output modality by indexing the corresponding vocabulary. Our subsequent ablation experiments also confirm that OmniMamba's dual-vocabulary design is one of the key factors for efficient training.

3.4 Task Specific LoRA

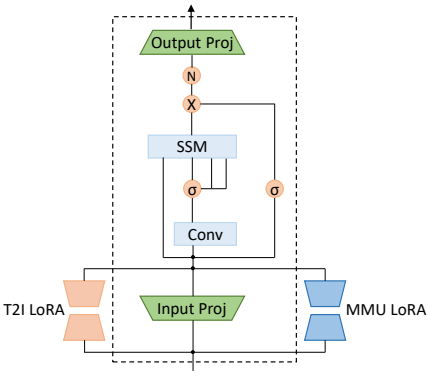

Fig. 3: The Mamba-2 block with task-specific LoRA.

To enhance the model's adaptability to specific tasks, we introduce task-specific adapters. We hypothesize that explicitly parameterizing the selection in SSMs based on task can enhance the data efficiency of multimodal training [15]. Specifically, to avoid introducing excessive parameters, we use LoRA [26] as the adapter. In OmniMamba, task-specific LoRA is applied only to the input projection of each Mamba-2 layer, as illustrated in Fig 3. When performing a specific task, the input linear projection and task-specific LoRA work together to effectively address the task. For instance, when the model performs a multimodal understanding (MMU) task, the MMU LoRA route is activated, while the text-to-image (T2I) LoRA route is dropped. Explicitly activating the corresponding adapter to assist in task execution helps improve data efficiency in training [15, 65].

3.5 Decoupled Training Strategy

We propose a decoupled two-stage training strategy to address data imbalance between understanding and generation tasks while improving training efficiency, as illustrated in Fig. 4. This approach consists of (1) a **Task-Specific Pre-Training** stage for module-specific initialization and modality alignment, and (2) a **Unified Fine-Tuning** stage for unified multi-task training.

Decoupling Rationale The first stage separates multimodal understanding (MMU) and text-to-image (T2I) generation tasks to prioritize modality alignment without data ratio constraints. Unlike joint pre-training methods (e.g., JanusFlow [47] with a fixed 50:50 MMU-T2I data ratio), our approach trains task-specific modules independently, enabling flexible dataset scaling (665K MMU vs. 83K T2I samples). Only randomly initialized components—linear projection and MMU LoRA for understanding, T2I LoRA and image head for generation—are trained, while the core Mamba-2 model remains frozen. This eliminates competition between tasks during early learning and allows asymmetric data utilization.

Stage 1: Task-Specific Pre-Training It contains: **MMU Pre-Training**: Trains the linear projection and MMU LoRA to align visual-textual representations. The T2I LoRA path is disabled to isolate understanding-task learning. **T2I Pre-Training**: Optimizes the T2I LoRA and image decoder for visual synthesis. The MMU LoRA path is disabled to focus on generation capabilities.

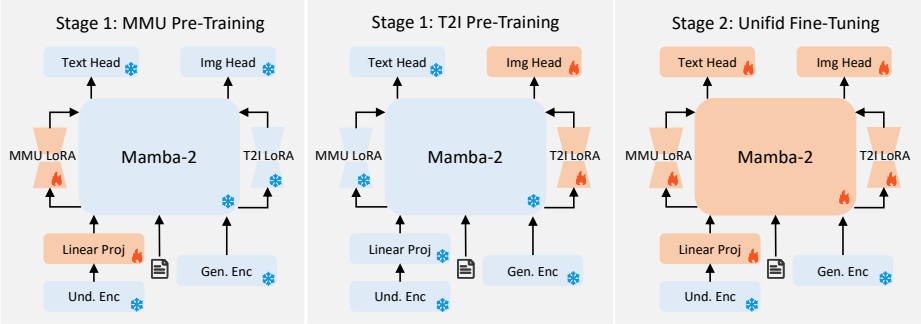

Fig. 4: Training strategy of OmniMamba. The trainable components are indicated by a flame symbol, while the frozen ones are represented by snowflakes. The dashed arrows indicate that this route is temporarily dropped and does not participate in model training.

Stage 2: Unified Fine-Tuning Inspired by multi-task frameworks [31, 81], we freeze the visual encoder and train all other modules while preserving task-specific LoRA independence. During each forward pass: (1) MMU and T2I computations use their respective LoRA branches; (2) Losses from both tasks are summed for a unified backward pass. This balances parameter sharing (via the frozen backbone) and task specialization (via isolated LoRA paths), enabling synergistic learning while mitigating interference between understanding and generation objectives.

3.6 Training Objective

Since OmniMamba uses the auto-regressive paradigm to handle both multimodal understanding and visual generation tasks, we only need to use the standard cross-entropy loss for next-token prediction during training.

4 Experiment

4.1 Data

To achieve the goal of data efficiency, we aim to train OmniMamba using as few high-quality image-text pairs as possible.

Multimodal Understanding Data In the first pretrain stage, the training data consists of 676K image-text pairs, all of which are sourced from publicly available datasets. These include 118K images from COCO [40] and 558K images from LLaVA-1.5 pre-training data [44].

In the second fine-tune stage, we also exclusively use publicly available datasets following Cobra [76].

1. The mixed dataset used in LLaVA-1.5 [44] consists of 665K visual multi-turn conversations.
2. LVIS-Instruct-4V [64], which comprises 220K images accompanied by visually aligned and context-aware instructions generated by GPT-4V.
3. LRV-Instruct [42], a large-scale visual instruction dataset of 400K samples, designed to mitigate hallucination issues across 16 vision-and-language tasks.

Visual Generation Data To facilitate better reproducibility and further exploration by the community, we use only 83K images from the MS-COCO 2014 dataset [40] for text-to-image generation training.

Overall, the training data for our unified multimodal generation model consists of fewer than 2 million image-text pairs. In contrast to previous works that rely on over 100M or even over 1B pairs, our approach is highly training efficient.

Data Formats Following Show-o [69], we use special tokens to unify the data formats for both multimodal understanding and visual generation tasks. The multimodal understanding data is structured as:

[MMU][SOI]{image tokens}[EOI][SOT]{text tokens}[EOT].

While the visual generation data is:

[T2I][SOT]{text tokens}[EOT][SOI]{image tokens}[EOI].

Specifically, [MMU] and [T2I] are pre-defined task tokens used to guide the model in performing the corresponding task. [SOT] and [EOT] are used to represent the beginning and end of text tokens, respectively. Similarly, [SOI] and [EOI] represent the beginning and end of image tokens.

4.2 Implementation Details

Our core model is based on Mamba-2-1.3B, which consists of 48 layers of Mamba-2 blocks. In our primary experiments, the input image resolution for the multimodal understanding task is 384, while the image resolution for the visual generation task is 256.

For multimodal understanding, we combine DINOv2 [49] and SigLIP [75] as the image encoder, while for visual generation, we use the VQVAE trained by LlamaGen [58] as the image encoder. We incorporate task-specific LoRA into the input projector of each Mamba-2 block and set the LoRA rank to 8, which results in only a 0.65% increase in parameters. All training stages use the AdamW optimizer with β_1 set to 0.9 and β_2 set to 0.95. We adopt cosine annealing with warm-up as the learning rate schedule. Weight decay is set to 0, and gradient clipping is applied with a threshold of 1.0. All training is conducted on NVIDIA A800 GPUs using BF16 precision.

Table 1: Comparison with other methods on multimodal understanding benchmarks. "Und. only" refers to models that only perform multimodal understanding task, while "Unified" refers to models that unify both multimodal understanding and visual generation tasks. Models with a similar number of parameters to ours are highlighted in light blue for emphasis.

Type	Model	LLM Params	Res.	POPE↑	MME-P↑	VQAv2$_{test}$↑	GQA↑	MMMU↑
Und. Only	LLaVA-Phi [81]	Phi-2-2.7B	336	85.0	1335.1	71.4	-	-
	LLaVA [45]	Vicuna-7B	224	76.3	809.6	-	-	-
	Emu3-Chat [66]	8B from scratch	512	85.2	-	75.1	60.3	31.6
	LLaVA-v1.5 [44]	Vicuna-13B	448	86.3	1500.1	81.8	64.7	-
	InstructBLIP [9]	Vicuna-13B	224	78.9	1212.8	-	49.5	-
	MobileVLM [7]	MobileLLaMA-1.4B	336	84.5	1196.2	-	56.1	-
	MobileVLM-V2 [8]	MobileLLaMA-1.4B	336	84.3	1302.8	-	59.3	-
	LLaVA-v1.5-Phi-1.5 [69]	Phi-1.5-1.3B	336	84.1	1128.0	75.3	56.5	30.7
Unified	LWM [43]	LLaMA2-7B	256	75.2	-	55.8	44.8	-
	Chameleon [60]	7B from scratch	512	-	-	-	-	22.4
	LaVIT [30]	7B from scratch	256	-	-	66.0	46.8	-
	Emu3 [66]	8B from scratch	512	85.2	1243.8	75.1	60.3	31.6
	Janus [68]	DeepSeek-LLM-1.3B	384	87.0	1338.0	77.3	59.1	30.5
	JanusFlow [47]	DeepSeek-LLM-1.3B	384	88.0	1333.1	79.8	60.3	29.3
	Show-o [69]	Phi-1.5-1.3B	512	80.0	1097.2	69.4	58.0	26.7
	OmniMamba	Mamba-2-1.3B	384	86.3	1290.6	77.7	60.8	30.6

4.3 Quantitative Results

Multimodal Understanding We evaluate OmniMamba's multimodal understanding capabilities on a wide range of vision-language benchmarks, including POPE [38], MME [18], GQA [29], MMMU [74]. The results are shown in Tab 1. Compared to models with a similar number of parameters, OmniMamba surpasses understanding-specific models such as LLaVA-v1.5-Phi-1.5 [69], MobileVLM [7], and MobileVLMv2 [8]. It also outperforms the unified understanding and generation model Show-o [69] and achieves competitive performance compared to the state-of-the-art unified model JanusFlow [47]. Notably, while Show-o utilizes 2B image-text pairs and JanusFlow leverages over 65M image-text pairs, OmniMamba achieves competitive performance by using only 2M image-text pairs for training.

Visual Generation We evaluate OmniMamba for text-to-image generation on the widely recognized MS-COCO [40] benchmark dataset. To quantify image quality, we report the FID score [24]. Consistent with previous literature, we randomly select 30K prompts from the MS-COCO validation set and generate corresponding images to compute the FID score. The results are shown in Tab 2, where our model achieves the best visual generation performance on the MS-COCO validation dataset. Notably, models such as Show-o [69] and PixArt [71] are trained on external large-scale datasets and further fine-tuned on COCO-like datasets (e.g., OpenImages [33]) before evaluating zero-shot on the MS-COCO validation set. In contrast, to avoid introducing excessive additional data, both our model and U-ViT [2] are trained solely on the MS-COCO training set and evaluated on the MS-COCO validation set.

Table 2: Compare visual generation capability with other methods on MS-COCO validation dataset.

Type	Model	Params	Images	FID-30K↓
Gen. Only	DALL·E [55]	12B	250M	27.5
	GLIDE [48]	5B	250M	12.24
	DALL·E 2 [54]	6.5B	650M	10.39
	SDv1.5 [56]	0.9B	2000M	9.62
	PixArt [4]	0.6B	25M	7.32
	Imagen [57]	7B	960M	7.27
	Parti [71]	20B	4.8B	7.23
	Re-Imagen [5]	2.5B	50M	6.88
	U-ViT [2]	45M	83k(coco)	5.95
Unified	CoDI [59]	-	400M	22.26
	SEED-X [21]	17B	-	14.99
	LWM [43]	7B	-	12.68
	DreamLLM [14]	7B	-	8.76
	Show-o [69]	1.3B	35M	9.24
	OmniMamba	1.3B	83k(coco)	**5.50**

4.4 Qualitative Results

We present qualitative evaluations of our OmniMamba for both multimodal understanding and visual generation tasks. Fig 5 showcases our model's capabilities in scene description and text-guided generation. Additional visualization results can be found in the Appendix.

4.5 Inference Speed and GPU Memory usage

We compared the generation speed and GPU memory usage of OmniMamba with other Transformer-based models in both multimodal understanding and visual generation tasks. All the evaluations were done on the same single NVIDIA 4090 GPU with FP16 precision.

Table 3: The Inference Speed in Multimodal Understanding Task.

Sequence Length	SmolLM2-1.7B	OmniMamba-1.3B
1024	56.94	**241.6**
2048	54.69	**245.1**
4096	57.49	**247.1**
8192	57.88	**247.5**
16384	57.40	**248.1**

In multimodal understanding task, all models received the same example image. We used the same prompt, "Please describe the image in detail." and

Table 4: Image Generation Speed in Visual Generation Task. OmniMamba achieves 7.0 × faster image generation speed compared to Show-o and 5.6 × faster compared to JanusFlow.

Model	Gen_{avg} (Image/s)	$Total$ (s)
Show-o [69]	0.81	19.66
JanusFlow [47]	1.02	15.64
OmniMamba	**5.68**	**2.81**

removed the token generation limit to test their generation speed. The results are shown in the Fig 1. OmniMamba demonstrates 119.2× speedup at a sequence length of 16k, and saves 63.0% GPU memory at a sequence length of 23k compared to Show-o-256. Meanwhile, at a sequence length of 100k, OmniMamba achieves a 10.2 × speedup and 40.4% GPU memory savings compared to JanusFlow-384, which is accelerated by FlashAttention-2 [10]. Notably, Show-o-256 indicates that the input image resolution is 256. Due to the design of its omni-attention mechanism being incompatible with FlashAttention-2, it was not used during testing. Similarly, JanusFlow-384 represents an input image resolution of 384, with FlashAttention-2 applied during testing for acceleration.

To further validate OmniMamba's efficiency, we benchmarked it on the MMU task against SmolLM2-1.7B [1], an auto-regressive Transformer model of comparable size. To ensure a rigorous test, we accelerated this Transformer baseline using FlashAttention-2 and kv cache. Tab. 3 shows the inference speed (tokens/sec) on a single NVIDIA 4090 GPU.

Similarly, in the text to image task, we used the same prompt, "A Picture". The models generate images with a batch size of 16 and a resolution of 256. As shown in Tab. 4, our model achieves image generation speeds that are 7.0 × faster than Show-o and 5.6 × faster than JanusFlow.

4.6 Ablation Studies

We conducted a series of ablation studies to verify the effectiveness of each design in OmniMamba. In this section, all ablation studies are conducted based with the understanding visual encoder replaced by CLIP [52], an input resolution of 224, and a reduced number of training steps (kept consistent across all ablation experiments), while keeping all other settings unchanged.

Impact of Decoupling Encoder for the Two Tasks Recognizing that generative encoders prioritize pixel reconstruction while understanding encoders are tailored for semantic comprehension, we employ decoupled encoders in OmniMamba. Ablation studies were conducted to validate this design, as shown in Exp1 and Exp2. These findings confirm that employing task-decoupled encoders is a critical factor for enabling efficient model training, especially when data is limited.

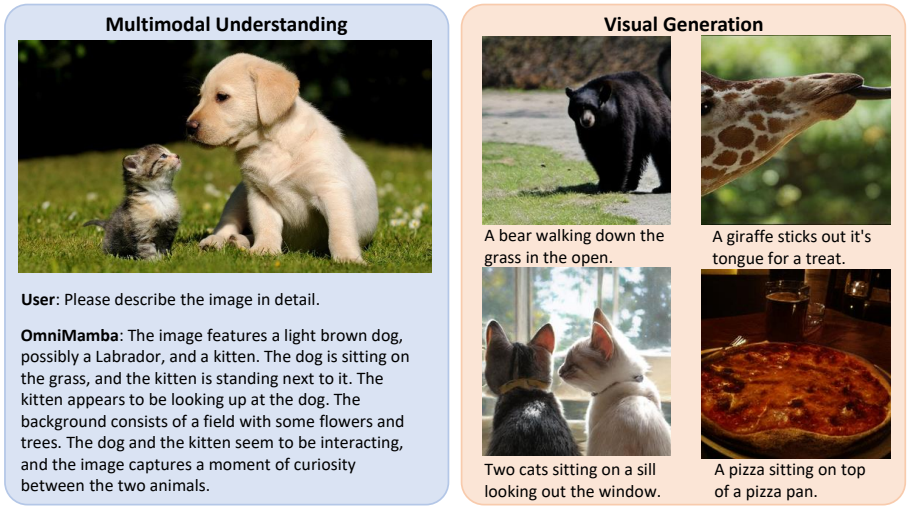

Fig. 5: Qualitative results of OmniMamba on multimodal understanding and visual generation.

Table 5: Ablation studies on architectural variants, encoder decoupling, vocabulary decoupling, task-specific adapters, and parameter scaling in OmniMamba.

#Exp	LLM Params	Decoupled Enc.	Decoupled Vocab	Task LoRA	POPE↑	MME↑	GQA↑	FID-30K↓
1	Mamba2-130M	✗	✓	✓	65.6	598	35.7	31.4
2	Mamba2-130M	✓	✓	✓	80.6	930	46.1	23.7
3	GPT-2-124M	✓	✓	✓	66.7	653	31.7	41.6
4	SmolLM2-135M	✓	✓	✓	81.4	1127	45.1	25.6
5	Mamba2-370M	✓	✗	✓	80.8	1036	53.6	19.1
6	Mamba2-370M	✓	✓	✗	81.2	1003	54.0	14.4
7	Mamba2-370M	✓	✓	✓	81.9	1100	55.3	10.3
8	Mamba2-1.3B	✓	✓	✓	86.3	1291	60.8	5.5

Impact of Decoupling Vocabulary for the Two Tasks To guide the model in generating specific modalities from a structural design perspective, we employ modality-decoupled vocabularies in the default OmniMamba. We conducted ablation studies on this design. As shown in Tab 5, Exp1 and Exp3 utilize a modality-unified vocabulary and modality-decoupled vocabularies, respectively. The results demonstrate that the decoupled vocabularies enable more efficient model training and yield better performance. Notably, when using the modality-unified vocabulary for visual generation, the model occasionally produces text-related tokens, which requires additional post-processing to ensure correct visual generation. This further indicates that the model needs additional training to effectively learn modality-specific generation.

Impact of Task Specific Adapter We conducted ablation studies on the introduced task-specific adapter module, and the results are shown in Tab 5. Exp6 and Exp7 represent the model without and with the task-specific adapter,

respectively. The experiments demonstrate that the task-specific adapter helps the model efficiently learn both multimodal understanding and visual generation with a minimal amount of training image-text pair data (2M).

Impact of LLM Parameter Scale To assess the scalability of our method, we conducted ablation studies on models of varying sizes. As shown in Exp2, Exp7 and Exp8 in Tab. 5 indicate that our model benefits from increased capacity, achieving better performance as the model size grows.

Impact of Architectural Variants To demonstrate the feasibility of our method across different architectures, we replace the LLM with GPT-2-124M [53] and SmolLM2-135M [1] while keeping the same settings. Exp2, Exp3 and Exp4 in Tab. 5 illustrate the adaptability of our method to different architectures, and Mamba2 outperforms Transformer with higher efficiency.

5 Conclusion

We presented OmniMamba, the first Mamba-2-based unified multimodal understanding and visual generation framework that achieves competitive performance with remarkable inference and training efficiency. By introducing three key innovations: decoupled vocabularies to disentangle modality-specific semantics, task-specific LoRA modules for parameter-efficient adaptation, and a two-stage decoupled training strategy to resolve data imbalance, OmniMamba achieves comparable performance with JanusFlow and even surpasses Show-o using only 2M image-text pairs for training. Moreover, OmniMamba exhibits outstanding inference efficiency. It achieves a 119.2× speedup with a sequence length of 16k and a 63% reduction in GPU memory at a sequence length of 23k, compared to Show-o. With a sequence length of 100k, it delivers a 10.2× speedup and saves 40.4% of GPU memory compared to JanusFlow. These results validate that our proposed OmniMamba is both training and inference efficient, with the potential to enable more ordinary researchers to participate in the wave of unified model innovation. However, due to the limited scale of training data, our model's performance remains slightly below SOTA methods. Exploring the trade-off between training data volume and model performance will be a key focus of our future work.

6 Limitation

Although our OmniMamba achieves promising results with a very small amount of data, the limited data volume constrains the model's performance. Furthermore, unlike previous works that leverage large-scale, high-quality datasets such as LAION-aesthetics, we rely solely on the MS-COCO dataset for visual generation. As a result, the quality of generated images, particularly for human faces, remains coarse. Exploring the trade-off between dataset scale and model performance will be a key focus of our future work.

Additionally, while Mamba-2 demonstrates exceptional inference efficiency, its foundational capabilities remain weaker compared to the extensively studied Transformer. Furthermore, Mamba-2 has only been trained on sequences of up to 2048 tokens, limiting its ability to handle ultra-long sequences and hindering its extension to advanced techniques such as Chain-of-Thought (CoT) [67] or reinforcement learning [23]. Enhancing Mamba-2's foundational capabilities and its capacity to model ultra-long sequences will be critical areas for future investigation.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (NSFC U25B2067).

References

1. Allal, L.B., Lozhkov, A., Bakouch, E., Blázquez, G.M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A.P., Srivastav, V., et al.: Smollm2: When smol goes big–data-centric training of a small language model. arXiv preprint arXiv:2502.02737 (2025)
2. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22669–22679 (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
5. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. arXiv preprint arXiv:2209.14491 (2022)
6. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
7. Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., et al.: Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886 (2023)
8. Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, X., Hu, Y., Lin, X., Zhang, B., et al.: Mobilevlm v2: Faster and stronger baseline for vision language model. arXiv preprint arXiv:2402.03766 (2024)
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
10. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)

11. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
14. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
15. Dou, S., Zhou, E., Liu, Y., Gao, S., Zhao, J., Shen, W., Zhou, Y., Xi, Z., Wang, X., Fan, X., et al.: Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. arXiv preprint arXiv:2312.09979 **4**(7) (2023)
16. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
17. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
18. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), https://arxiv.org/abs/2306.13394
19. Ge, Y., Ge, Y., Zeng, Z., Wang, X., Shan, Y.: Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041 (2023)
20. Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. arXiv preprint arXiv:2310.01218 (2023)
21. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
22. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
23. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
27. Hu, V.T., Baumann, S.A., Gui, M., Grebenkova, O., Ma, P., Fischer, J., Ommer, B.: Zigma: A dit-style zigzag mamba diffusion model. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
28. Huang, W., Pan, J., Tang, J., Ding, Y., Xing, Y., Wang, Y., Wang, Z., Hu, J.: Ml-mamba: Efficient multi-modal large language model utilizing mamba-2. arXiv preprint arXiv:2407.19832 (2024)

29. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
30. Jin, Y., Xu, K., Xu, K., Chen, L., Liao, C., Tan, J., Huang, Q., Chen, B., Lei, C., Liu, A., et al.: Unified language-vision pretraining in llm with dynamic discrete visual tokenization. arxiv 2024. arXiv preprint arXiv:2309.04669
31. Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., Sadigh, D.: Prismatic vlms: Investigating the design space of visually-conditioned language models. arXiv preprint arXiv:2402.07865 (2024)
32. Koh, J.Y., Fried, D., Salakhutdinov, R.R.: Generating images with multimodal language models. Advances in Neural Information Processing Systems **36** (2024)
33. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
34. LAION: Laion-coco 600m. https://laion.ai/blog/laion-coco (2022)
35. Li, H., Yang, J., Wang, K., Qiu, X., Chou, Y., Li, X., Li, G.: Scalable autoregressive image generation with mamba. arXiv preprint arXiv:2408.12245 (2024)
36. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
37. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
38. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
39. Liao, B., Tao, H., Zhang, Q., Cheng, T., Li, Y., Yin, H., Liu, W., Wang, X.: Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation. arXiv preprint arXiv:2502.13145 (2025)
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
41. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
42. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2023)
43. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with ringattention. arXiv e-prints pp. arXiv–2402 (2024)
44. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
45. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
46. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)
47. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for uni-

fied multimodal understanding and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7739–7751 (2025)
48. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
50. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
51. Qiao, Y., Yu, Z., Guo, L., Chen, S., Zhao, Z., Sun, M., Wu, Q., Liu, J.: Vl-mamba: Exploring state space models for multimodal learning. arXiv preprint arXiv:2403.13600 (2024)
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
53. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
54. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
55. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
57. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
58. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
59. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems **36** (2024)
60. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
61. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
62. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
63. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
64. Wang, J., Meng, L., Weng, Z., He, B., Wu, Z., Jiang, Y.G.: To see is to believe: Prompting gpt-4v for better visual instruction tuning. arXiv preprint arXiv:2311.07574 (2023)

65. Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., Zhou, C.: One-peace: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172 (2023)
66. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
67. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
68. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
69. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: International Conference on Learning Representations. vol. 2025, pp. 28240–28264 (2025)
70. Yang, S., Wang, B., Shen, Y., Panda, R., Kim, Y.: Gated linear attention transformers with hardware-efficient training. arXiv preprint arXiv:2312.06635 (2023)
71. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
72. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023)
73. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. arXiv preprint arXiv:2411.00776 (2024)
74. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
75. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
76. Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S., Wang, D.: Cobra: Extending mamba to multi-modal large language model for efficient inference. arXiv preprint arXiv:2403.14520 (2024)
77. Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., Liu, X., Wu, J., Huang, L.: Tinyllava: A framework of small-scale large multimodal models. arXiv preprint arXiv:2402.14289 (2024)
78. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. In: International Conference on Learning Representations. vol. 2025, pp. 6446–6469 (2025)
79. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
80. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: International Conference on Machine Learning. pp. 62429–62442. PMLR (2024)

81. Zhu, Y., Zhu, M., Liu, N., Xu, Z., Peng, Y.: Llava-phi: Efficient multi-modal assistant with small language model. In: Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited. pp. 18–22 (2024)