MANGO: Unleashing Image Generation Capability of Unified Multimodal Models

Yi Wang[1*], Mushui Liu[5*], Wanggui He[5*], Hanyang Yuan[1*], Ziwei Huang[2], Guanghao Zhang[5], Wenkai Fang[1], Haoze Jiang[2], Shengxuming Zhang[2], Weilong Dai[5], Haofei Zhang[3,4], Mingli Song[1,2,3,4], Hao Jiang[5], and Jie Song[2✉]

[1] College of Computer Science and Technology, Zhejiang University
[2] School of Software Technology, Zhejiang University
[3] State Key Laboratory of Blockchain and Data Security, Zhejiang University
[4] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[5] Alibaba Group
{y_w,lms,sjie}@zju.edu.cn, {wanggui.hwg,aoshu.jh}@taobao.com
[*] Equal contribution. [✉] Corresponding author.

Abstract. Unified multimodal models have demonstrated remarkable versatility in multimodal understanding and generation. However, they struggle with image generation under complex compositional instructions, a setting frequently encountered in real-world applications. To address this, we present MANGO, a novel model that unleashes image generation capability at both its performance floor and ceiling. (1) To raise the performance floor, we introduce Functionality-Oriented Transformers, an architectural design that assigns each Transformer branch to a dedicated function, such as image generation or image understanding. This function-disentangled design alleviates cross-functional interference inherent in mainstream modality-oriented architectures, thereby unleashing the model's image generation capacity. (2) To elevate the performance ceiling, we propose Paint-CoT, a framework that enables image generation through a human-like artistic chain of thought, consisting of planning, acting, reflection, and correction. Leveraging the model's multimodal understanding and reasoning capabilities, this reasoning-guided process further unleashes image generation capacity. Further, to overcome the difficulty and high cost of collecting multi-step aligned chain-of-thought data for end-to-end training, we develop a multi-task joint training paradigm that decomposes the full reasoning-generation pipeline into manageable subtasks. This strategy maximizes the utilization of existing or readily constructed supervision, enabling effective Paint-CoT-based image generation without requiring fully aligned multi-step annotations. Extensive experiments demonstrate that MANGO consistently achieves superior performance across diverse image generation benchmarks, delivering substantial improvements in complex image generation scenarios.

Keywords: Unified Multimodal Models · Chain of Thought

1 Introduction

Unified multimodal models, such as GPT-5 [57] and Gemini 3 [24], aim to integrate multimodal understanding and generation within a single model. Such an all-

in-one paradigm is widely regarded as an ideal future model form, as it enables more elegant and efficient scaling, training, and deployment compared to complex systems composed of multiple specialized expert models. Early works [37,51,67] adopt a next-token prediction paradigm that represents all modalities as discrete tokens. However, this discrete tokenization strategy does not align well with the continuous nature of visual signals, such as images and videos, thereby constraining their generative potential. Subsequent studies [53,83,90] explore a hybrid generative paradigm that predicts text autoregressively as discrete tokens while generating images as continuous signals via diffusion processes. However, these models rely on a single Transformer branch to process all modalities. Sharing a unified branch across inherently heterogeneous modalities may introduce implicit modality conflicts and limit specialization. To mitigate this issue, MoT [40] and LlamaFusion [64] introduce modality-oriented architectures, in which the model backbone consists of multiple Transformer branches, each dedicated to processing a single modality, e.g., text or image. These designs enable unified multimodal models to achieve strong versatility in multimodal understanding and generation.

While unified multimodal models can achieve comparable or even superior image generation performance to dedicated text-to-image diffusion models [3,16,58], they struggle under complex compositional instructions, such as multi-object co-occurrence, attribute binding, and spatial constraints. These scenarios frequently arise in real-world applications and pose a vital challenge. As illustrated in the first two rows of Fig. 1, unified multimodal models exhibit several limitations, including concept confusion in multi-object scenarios, failures even with only two objects, attribute misalignment (e.g., incorrect color binding), spatial inconsistencies (e.g., improper positioning), and structural defects such as incomplete object parts. To address this, we present MANGO, a novel model that unleashes image generation capability at both its performance floor and ceiling.

To raise the performance floor, we first propose Functionality-Oriented Transformers (FoT), an architectural design to empower complex image generation. We identify that even though mainstream modality-oriented architectures [40,64], disentangle modality conflicts, they still suffer from potential functional conflicts that hinder image generation capability. Specifically, image understanding and image generation are processed by a shared visual modal branch. However, these two capabilities pursue fundamentally different objectives: image understanding is optimized under next-token prediction to align visual features with textual semantics, whereas image generation is trained to minimize diffusion denoising loss, driving visual features toward accurate pixel-level noise trajectory prediction. To address this, FoT (as illustrated in Fig. 2) assigns each Transformer branch to a dedicated function, such as image understanding or image generation. By disentangling functionalities, FoT alleviates cross-functional interference, thereby unleashing the model's image generation capacity and improving its performance floor for complex image generation scenarios.

To elevate the performance ceiling, we introduce Paint-CoT, a human-like artistic Chain-of-Thought (CoT) framework to further push the boundaries of complex image generation. While recent works [15,18,31] extend CoT reasoning

Fig. 1: Limitations of current unified multimodal models in complex compositional image generation scenarios. The results generated by JanusFlow and Show-o exhibit noticeable confusion and defects. In contrast, MANGO effectively mitigates these issues.

to this domain, they still face certain constraints: GoT [18] adopts a cascaded system design, where a Multimodal Large Language Model (MLLM) performs reasoning while a separate diffusion model handles image generation. This design is incompatible with our unified single-model setting. Other approaches [15, 31] rely on complex reinforcement learning with carefully designed reward functions to enable content planning before image generation. While planning improves generation quality, it remains insufficient for complex compositional scenarios. Models with planning may fail to accurately render all visual components in a single generation attempt, leaving residual defects, as illustrated in Fig. 3. To overcome these constraints, Paint-CoT incorporates a reflection–correction mechanism, thereby mitigating residual image defects and further elevating the performance ceiling of image generation in complex scenarios.

Furthermore, collecting fully aligned multi-step chain-of-thought data for end-to-end training is difficult and costly. As existing datasets typically provide only prompt–image pairs, constructing full multi-step supervision would require expensive annotation and precise cross-step alignment. To transcend this limitation, we develop a multi-task joint training paradigm that decomposes the full Paint-CoT pipeline into manageable subtasks, maximizing the utilization of existing datasets and enabling effective Paint-CoT-based image generation without requiring fully aligned multi-step annotations.

In summary, our main contributions are as follows:

- We propose MANGO, a novel model that unleashes both the performance floor and ceiling of image generation, achieving strong results on diverse image generation benchmarks and complex compositional scenarios.
- We introduce FoT, a functionality-oriented architectural design that disentangles functional conflicts between image understanding and image generation, improving the model's image generation performance floor.

– We propose Paint-CoT, along with a multi-task joint training paradigm, enabling image generation with reasoning guidance through a human-like artistic CoT while avoiding the need for fully aligned multi-step supervision, thereby elevating the performance ceiling for complex image generation.

2 Related Works

2.1 Unified Multimodal Models

GPT series [56] in the language domain have inspired the development of multimodal large language models (MLLMs) [2,8,46], which extend language models to the vision domain by integrating vision and language understanding. However, these models still lack image generation capability. To address this, several works couple large language models (LLMs) with text-to-image diffusion models [21,65,78]. While straightforward, such systems, composed of multiple specialized expert models, are difficult to scale, train, and deploy efficiently. Subsequent studies explore unified multimodal modeling by representing both text and images as discrete tokens [37,51,67], though this discrete tokenization often leads to limited image generation quality. Hybrid generative architectures [53,83,91] further combine autoregressive text generation with diffusion-based image generation within a unified model. However, these models typically rely on a shared Transformer branch, which may introduce modality conflicts. To alleviate this issue, MoT [40] and LlamaFusion [64] propose modality-oriented architectures, where the model backbone consists of multiple Transformer branches, each dedicated to processing a specific modality. Nevertheless, potential functional conflicts remain, as image understanding and image generation are still handled by the same visual branch. Inspired by MoT, BAGEL [13] assigns separate branches for understanding and generation, respectively, but overlooks modality conflicts, such as text understanding and image understanding being processed within the same branch. To address these, we introduce FoT, which preserves complete text–image modality separation while alleviating cross-functional interference between image understanding and image generation, thereby raising the model's image generation performance floor. Specifically, FoT consists of three branches: (1) Linguistic Branch for text understanding and generation, as these functions are naturally aligned in language modeling; (2) Semantic Visual Branch for image understanding; and (3) Generative Visual Branch for image generation.

2.2 Chain of Thought

CoT [42,76,92] is validated to be effective in improving the ability of LLMs to solve complex tasks via step-by-step reasoning, and has increasingly been adopted in other domains, including MLLMs and vision-language-action (VLA) models. Existing CoT approaches can be broadly categorized into two types: (1) learned entirely end-to-end, where reasoning ability emerges directly from large-scale training, like OpenAI o1 [42] and DeepSeek-R1 [26], which cannot be adopted here due to the lack of multi-step data for end-to-end CoT training of image generation; and (2) human-defined for key steps, where complex tasks are decomposed into human-defined key subproblems, as in [27,32,54,84,89], across

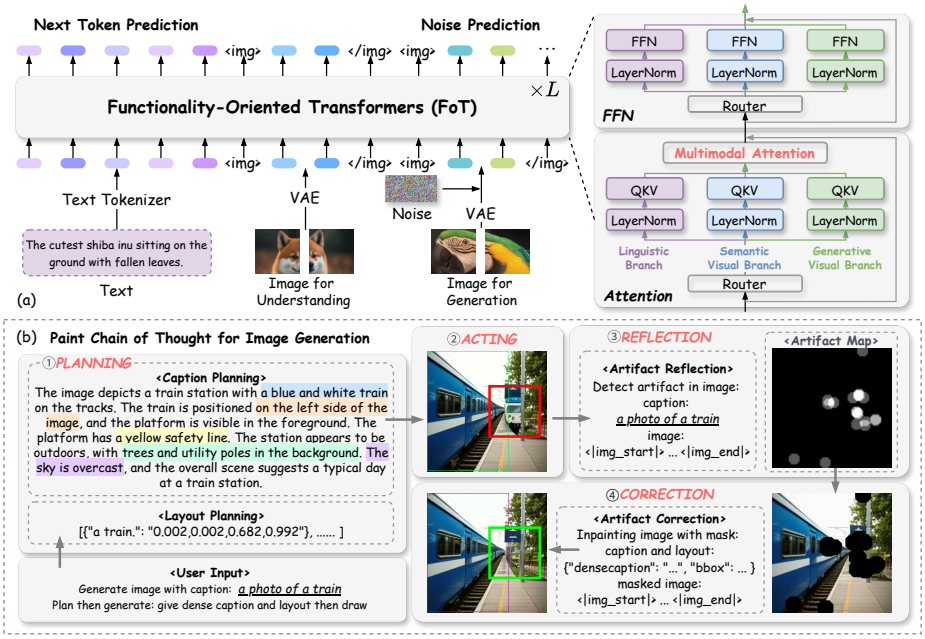

Fig. 2: Overall framework for MANGO. (a) The model architecture highlights FoT as the core component. (b) The illustration of Paint-CoT for image generation.

LLMs, MLLMs, and VLA models, which are mainly designed for specific tasks such as understanding and are not compatible with image generation scenarios. Recent works [15,18,31] extend CoT reasoning to the image generation domain. GoT [18] adopts a cascaded system design, where an MLLM like Qwen2.5-VL [2] performs reasoning while a separate diffusion model like SD-XL [58] handles image generation. This design is incompatible with our unified single-model setting. Other approaches [15,31] rely on complex reinforcement learning with carefully designed reward functions to enable content planning before image generation. While planning improves generation quality, it remains insufficient for complex compositional scenarios. Models with planning may fail to accurately render all visual components in a single generation attempt, leaving residual defects, as illustrated in Fig. 3. To address this, we propose Paint-CoT, a framework that enables image generation through a human-like artistic chain of thought consisting of planning, acting, reflection, and correction. By incorporating a reflection–correction mechanism, Paint-CoT mitigates residual image defects and further elevates the performance ceiling of image generation in complex scenarios, as empirically validated in Sec. 4.3.

3 Method

3.1 Preliminaries

We adopt a hybrid generative paradigm following MoT that predicts text autoregressively as discrete tokens while generating images as continuous signals via

diffusion. Below, we provide a brief review of these two modeling paradigms and offer a straightforward comparison to illustrate the potential functional conflicts arising from these two objectives when sharing a single Transformer branch.

Language Modeling. Let $z = (z_1, \ldots, z_N) \in V^N$ denote a sequence of discrete tokens drawn from a vocabulary V. Autoregressive language models factorize the joint distribution of these tokens using a causal decomposition:

$$P(z) = \prod_{i=1}^{N} P_\theta(z_i \mid z_{<i}), \tag{1}$$

where θ parameterizes the conditional distributions, enabling the model to predict the next token given all previous generated ones. Training of the language model aims to minimize the negative log-likelihood over the dataset:

$$\mathcal{L}_{\text{LM}} = \mathbb{E}_{z\sim\mathcal{D}} \left[-\sum_{i=1}^{N} \log P_\theta(z_i \mid z_{<i}) \right]. \tag{2}$$

For multimodal data, such as image-caption pairs (x, c), the model conditions each token not only on previous tokens $z_{<i}$ but also on visual features extracted from the image: $P_\theta(z_i \mid z_{<i}, \phi(x))$, where $\phi(\cdot)$ denotes the visual encoder.

Diffusion Modeling. We adopt the widely used Rectified Flow (RF) [17] for diffusion modeling, which defines a deterministic generative process for image data. Given images $\mathbf{x} \sim \mathcal{D}$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, we construct linear trajectories between noise and data:

$$\mathbf{x}_t = t\mathbf{x} + (1-t)\boldsymbol{\epsilon}, \quad t \in [0, 1]. \tag{3}$$

The velocity field model v_θ predicts straightening directions through:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t\sim U(0,1),\mathbf{x},\boldsymbol{\epsilon},c} \left[\|(\mathbf{x} - \boldsymbol{\epsilon}) - v_\theta(\mathbf{x}_t, t, c)\|_2^2 \right], \tag{4}$$

Based on this, it is obvious that image understanding and image generation pursue fundamentally different objectives. Image understanding is optimized via next-token prediction (Eq. (2)) to align visual features with textual semantics. In contrast, image generation is trained to minimize the diffusion denoising loss (Eq. (4)), guiding visual features toward accurate prediction of pixel-level noise trajectories. These objectives drive the shared visual features in the same branch toward substantially different representation targets, which may inevitably induce functional conflicts between image understanding and generation.

3.2 Input Representation

Text inputs $\boldsymbol{T}$ are transformed into an embedding sequence $\boldsymbol{x}_{\text{text}} \in \mathbb{R}^{L\times d}$ using Qwen2's tokenizer [1], where $\boldsymbol{L}$ is the sequence length and $\boldsymbol{d}$ is the embedding

dimension. An image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, with height $\boldsymbol{H}$ and width $\boldsymbol{W}$, is encoded into a latent representation using SD3's Variational Autoencoder (VAE) [17]. Following the approach from Transfusion [90], we compress 2×2 patches into a single vector, resulting in image tokens $\boldsymbol{x}_{\text{image}} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times d}$ after linear projection, where each token corresponds to a 16×16 pixel patch of the original image. The VAE is also used for image understanding tasks, consistent with prior works (TransFusion [90], LlamaFusion [64]). It preserves fine-grained, structurally rich visual features to facilitate detailed perception, which may benefit the reflection step of Paint-CoT. To accommodate and differentiate the newly introduced image representations, we have expanded the original vocabulary of Qwen2, which consists of 151,936 entities, by incorporating 6 functional special tokens.

3.3 Functionality-Oriented Transformers

We introduce FoT, an architectural design that assigns each Transformer branch to a dedicated function, such as image understanding or image generation. By disentangling functionalities, FoT preserves complete text–image modality separation while mitigating cross-functional interference between image understanding and generation, thereby unleashing the model's image generation capacity and improving its performance floor for complex image generation scenarios. Additionally, our design motivation is also supported by prior work [88], which demonstrates emergent modularity in pre-trained Transformers, where neurons naturally cluster into functionally specialized modules.

Specifically, FoT maintains a unified text branch, Linguistic Branch, since text understanding and generation share identical optimization objectives. For vision, it employs two separate branches: Semantic Visual Branch for image understanding and Generative Visual Branch for image generation, each optimized with distinct objectives. Unless otherwise specified, the Transformer architecture of all branches follows that of Qwen2-0.5B, with independent weights for all non-embedding parameters—including projection matrices in attention modules, feed-forward networks, and layer normalization—while sharing the multimodal routing and global multimodal attention modules, as illustrated in Fig. 2.

3.4 Forward Process

Denote the input as $\boldsymbol{x} = \boldsymbol{x}_T \oplus \boldsymbol{x}_C \oplus \boldsymbol{x}_N$, where $\boldsymbol{x}_T$, $\boldsymbol{x}_C$, $\boldsymbol{x}_N$ correspond to text tokens $\boldsymbol{x}_{text}$, clean image tokens $\boldsymbol{x}_{clean}$, and noise tokens $\boldsymbol{x}_{noised}$, respectively. The forward pass of a single FoT layer can then be expressed as:

$$\begin{aligned} \hat{\boldsymbol{x}}_i &= W_i(Router(LN(\boldsymbol{x}))), && i \in \{T, C, N\} \\ \hat{\boldsymbol{x}}_i^{q,k,v} &= W_i^{Q,K,V}(\hat{\boldsymbol{x}}_i), && i \in \{T, C, N\} \\ \hat{\boldsymbol{x}}^{rep} &= \hat{\boldsymbol{x}}_T^{rep} \oplus \hat{\boldsymbol{x}}_C^{rep} \oplus \hat{\boldsymbol{x}}_N^{rep}, && rep \in \{q, k, v\} \\ \hat{\boldsymbol{x}} &= Attn(\hat{\boldsymbol{x}}^q, \hat{\boldsymbol{x}}^k, \hat{\boldsymbol{x}}^v) + \boldsymbol{x}, \end{aligned} \tag{5}$$

where $\oplus$ denotes concatenation; T, C, N indicate Linguistic, Semantic Visual, and Generative Visual Branch, respectively. *Attn* denotes the Multimodal Attention

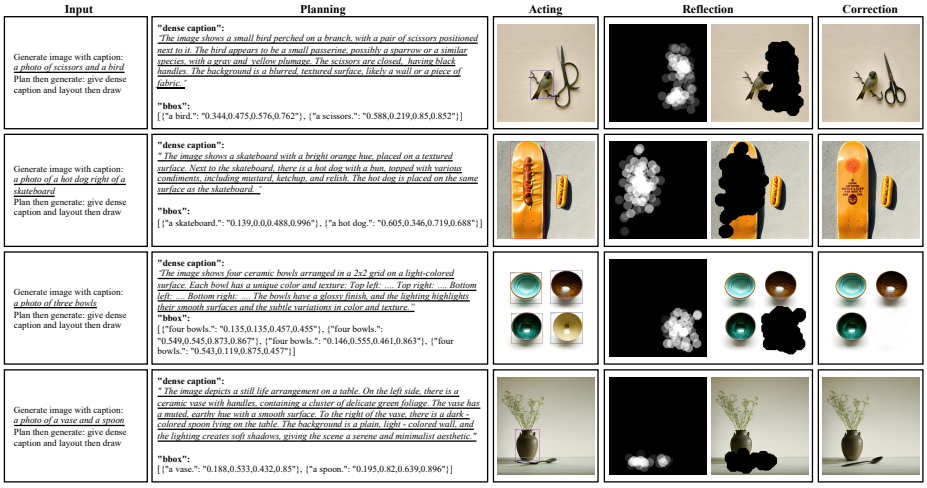

Fig. 3: Examples of Paint-CoT process, with each row corresponding to a different type of image defect corrected via reflection–correction. From top to bottom: Structural Incompleteness, Object Entanglement, Object Redundancy, and Object Distortion.

module. While branch parameters are decoupled, outputs interact and align at each layer through multimodal attention like [40, 64]. Locally, language tokens use causal attention, whereas vision tokens (clean and noisy) adopt bidirectional attention. Globally, all tokens follow a causal sequence, ensuring efficient loss and gradient computation without future information leakage.

The corresponding FFN module is simplified as:

$$\begin{aligned} \hat{\boldsymbol{x}}_i &= FFN_i(Router(LN(\boldsymbol{x}))), \quad i \in \{T, C, N\} \\ \hat{\boldsymbol{x}} &= \hat{\boldsymbol{x}}_T \oplus \hat{\boldsymbol{x}}_C \oplus \hat{\boldsymbol{x}}_N. \end{aligned} \tag{6}$$

3.5 Paint Chain of Thought for Image Generation

We propose the Paint-CoT, inspired by the human artistic workflow, where an artist first sketches object positions and outlines details, then reflects on and adjusts specific regions to achieve perfection. Specifically, Paint-CoT is defined as a sequence of explicit key reasoning steps, Planning, Acting, Reflection, and Correction, following the human-defined key steps manner discussed in Sec. 2.

Planning. Emulating an artist sketching object positions and outlining details before painting, the planning step—as illustrated in Paint-CoT process results in Fig. 3—consists of two components: detailed caption planning and layout box planning. Detailed caption planning drafts a more comprehensive and precise image caption without distorting the meaning of the input prompt. Layout box planning assigns each object in the input prompt to a reasonable relative position in a bounding-box manner, learned from large-scale image layouts.

Acting. In the acting step, MANGO generates images guided by the dense captions and layout boxes produced during the planning stage. We observe that using dense captions improves image fidelity, consistent with prior works [28,34] indicating that longer and more detailed prompts often yield better image quality. Furthermore, the layout boxes enable our model to place objects accurately within the specified areas. Additional examples demonstrating improvements in fidelity and adherence to layout are provided in Fig. 5 in Appendix B.

Reflection. Despite the planning and acting steps improving image generation, the model may still fail to render all parts of an image perfectly in a single attempt. Four main types of image defects that can be corrected through the reflection and correction steps are illustrated in Fig. 3. In the reflection step, the model uses the generated image and the input prompt as conditions to identify regions with defects, such as low aesthetic quality or misalignment with the prompt. This process generates an artifact heatmap, where higher confidence scores indicate areas requiring more substantial corrections.

Correction. In the correction step, our model integrates the artifact reflection map and planning rationale into the generation process, as shown in Fig. 2. Using this information, the model performs targeted inpainting to refine masked regions based on the artifact heatmap, thereby mitigating residual image defects and further elevating the performance ceiling of image generation in complex scenarios. Additional correction examples are provided in Fig. 3.

3.6 Multi-task Joint Training Paradigm

We develop a multi-task joint training paradigm that equips MANGO with Paint-CoT in a disentangled manner, overcoming the challenge of collecting multi-step aligned data. As illustrated in Fig. 2, a complete tuple includes {input prompt, detailed planning caption, layout planning boxes, first "wrong" image, artifact map, final "correct" image}. While available datasets provide only prompt-image pairs, we use MLLMs and detection models to generate planning captions and layout boxes. However, it is challenging to construct the first "wrong" image that simultaneously aligns with the prompt, planning, and final image while containing diverse realistic errors for subsequent reflection and correction steps. Moreover, using such a "wrong" image in end-to-end training would force the model to generate an incorrect image after planning, which contradicts our objective.

Specifically, we decouple the end-to-end training process into multiple tasks based on Paint-CoT steps: Planning and Acting Task, Reflection Task, and Correction Task. This approach allows the model to be trained with more readily available data for each subtask, while avoiding the drawback of forcing the model to generate erroneous images after planning in end-to-end training. Before Paint-CoT training, MANGO is pretrained to obtain basic text-to-image (T2I) and image-to-text (I2T) capabilities, with details provided in Appendix C.

For Planning and Acting Task, our model is trained with quadruples consisting of {input prompt, detailed caption, layout box, image}, expanded from prompt-image data pairs, where the model takes the input prompt and generates the detailed caption, layout box, and image. For Reflection Task, the model is trained to take the first generated image and input prompt to generate an artifact map, with artifact map labels produced from existing datasets and manually annotated data. For Correction Task, the model is trained in a standard inpainting style, refining masked regions based on all previous results such as detailed caption planning and the first generated image. Additional training details and data construction methods are provided in Appendix C and Appendix D.

4 Experiments

We employ AdamW [50] as the optimizer with parameter settings of $\beta_1 = 0.9, \beta_2 = 0.999$, and a weight decay of 0.02. The learning rate is configured to a constant value of 5×10^{-5}, incorporating a linear warm-up phase over 10,000 steps. The training utilized DeepSpeed's ZeRO-2 [61] optimization.

4.1 Main Results

GenEval Benchmarks. We evaluate on GenEval [23], a compositional image benchmark for assessing generative models under complex conditions. To further validate the scalability of our architecture, we also propose MANGO-4B, which replaces the default Qwen2-0.5B single-branch backbone with the larger Qwen2-1.5B model. As shown in Tab. 1, MANGO-4B achieves an overall score of 0.82, delivering superior or comparable performance to larger diffusion models like SD3 (0.68), unified multimodal models like Janus-Pro (0.80) and BAGEL (0.82), and CoT-guided models like GoT-R1 (0.75) and GoT (0.64). Despite its compact 1.3B parameter size, MANGO still outperforms significantly larger diffusion models, yielding a 9% improvement over SD 3 and a 10% gain over DALL·E 3 in overall score. Notably, it achieves remarkable 20% and 19% improvements over these two models on the challenging Attribute Binding metric, respectively. Furthermore, achieving an overall score of 0.77, the 1.3B MANGO outperforms both GoT (0.64) and GoT-R1 (0.75) despite their reliance on larger base models. It surpasses or performs comparably to these two CoT-guided models across almost all sub-metrics, firmly validating the superiority of our proposed approach. Additional qualitative comparisons with current models, including SD-XL, SD3, Show-o, and JanusFlow, are provided in Appendix B.

T2I-CompBench Benchmarks. T2I-CompBench [30] is also a compositional image benchmark used to further validate our model's generalization in complex scenarios. As shown in Tab. 2, MANGO achieves state-of-the-art scores in Spatial (35.71), Non-Spatial (34.19), and Complex (42.78), demonstrating strong performance in spatial arrangement, semantic consistency, and compositional reasoning for complex image generation. Furthermore, MANGO-4B consistently improves

Table 1: Evaluation of text-to-image generation on GenEval. * denotes models consisting of an MLLM and a diffusion model.

Model	Params.	Overall↑	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Text-to-Image Diffusion Models								
SD v1.5 [63]	1B	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD v2.1 [63]	1.3B	0.50	0.98	0.51	0.44	0.85	0.07	0.17
SD-XL [58]	3.4B	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SD 3 [16]	12.7B	0.68	0.98	0.84	0.66	0.74	0.40	0.43
DALL-E 2 [62]	4.5B	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALL-E 3 [3]	–	0.67	0.96	0.87	0.47	0.83	0.43	0.45
IF-XL [12]	10.1B	0.61	0.97	0.74	0.66	0.81	0.13	0.35
Unified Multimodal Models								
Chameleon [68]	34B	0.39	–	–	–	–	–	–
Transfusion [91]	7.3B	0.63	–	–	–	–	–	–
LWM [44]	7B	0.47	0.93	0.41	0.46	0.79	0.09	0.15
Janus-Pro [7]	7B	<u>0.80</u>	0.99	0.89	0.59	0.90	0.79	0.66
BAGEL [13]	7B	**0.82**	0.99	0.94	0.81	0.88	0.64	0.63
SEED-X* [22]	17B	0.49	0.97	0.58	0.26	0.80	0.19	0.14
GoT* [18]	2.8+3B	0.64	0.99	0.69	0.67	0.85	0.34	0.27
GoT-R1 [15]	7B	0.75	0.99	0.94	0.50	0.90	0.46	0.68
Show-o [81]	1.3B	0.53	0.95	0.52	0.49	0.82	0.11	0.28
Janus [77]	1.3B	0.61	0.97	0.68	0.30	0.84	0.46	0.42
JanusFlow [53]	1.3B	0.63	0.97	0.59	0.45	0.83	0.53	0.42
MANGO (Ours)	1.3B	0.77	0.99	0.86	0.71	0.82	0.60	0.64
MANGO-4B (Ours)	4.2B	**0.82**	0.99	0.89	0.76	0.87	0.68	0.70

Table 2: Evaluation of text-to-image generation on T2I-CompBench.

Model	Params.	Color↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	Complex↑
Text-to-Image Diffusion Models							
SD v1.5 [63]	1B	37.65	35.76	41.56	12.46	30.79	30.80
SD-XL [58]	3.4B	63.69	54.08	56.37	20.32	31.10	40.91
SD 3 [16]	12.7B	81.32	58.85	73.34	32.00	31.40	37.71
GORS [30]	-	66.03	47.85	62.87	18.15	31.93	33.28
PixArt-α [6]	0.6B	68.86	55.82	70.44	20.82	31.79	41.17
DALL-E 2 [62]	4.5B	57.50	54.64	63.74	12.83	30.43	36.96
Unified Multimodal Models							
Emu 3 [75]	7B	75.44	57.06	71.64	-	-	-
Janus-Pro [7]	7B	63.59	35.28	49.36	20.61	30.85	35.59
T2I-R1 [31]	7B	81.30	58.52	72.43	33.78	30.90	39.93
GoT-R1 [15]	7B	81.39	55.49	73.39	33.06	31.69	39.44
Show-o [82]	1.3B	56	41	46	20	30	29
MANGO (Ours)	1.3B	<u>82.37</u>	<u>59.81</u>	<u>74.21</u>	<u>35.71</u>	<u>34.19</u>	<u>42.78</u>
MANGO-4B (Ours)	4.2B	**86.63**	**61.92**	**78.75**	**40.45**	**37.41**	**45.29**

upon the 1.3B version across all metrics, further validating the effectiveness of our scaling strategy. Notably, larger models such as Janus-Pro and Emu 3 still lag behind our performance. Moreover, CoT-based models with larger 7B baseline models, such as T2I-R1 and GoT-R1, also underperform our approach on T2I-CompBench, further demonstrating the effectiveness of our approach.

WiSE Benchmarks. We further evaluate our model on the WISE benchmark [55], a representative and widely adopted reasoning-related image generation benchmark. The results in Tab. 3 demonstrate that our MANGO-4B outperforms

Table 3: Evaluation of Image Generation on WISE Benchmark.

Model	Params.	Cultural	Time	Space	Biology	Physics	Chemistry	Overall↑
Text-to-Image Diffusion Models								
SD v1.5 [63]	1B	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SD-XL [58]	3.4B	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD 3 [16]	12.7B	0.42	0.44	0.48	0.39	0.47	0.29	0.42
PixArt-Alpha [6]	0.6B	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Unified Multimodal Models								
Emu 3 [75]	7B	0.34	0.45	0.48	0.41	0.45	0.27	0.39
BAGEL [13]	7B	0.44	0.55	0.68	0.44	0.60	0.39	0.52
Janus-Pro [7]	7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
T2I-R1 [31]	7B	0.56	0.55	0.63	0.54	0.55	0.30	<u>0.54</u>
Show-o [82]	1.3B	0.28	0.36	0.40	0.23	0.33	0.22	0.30
Janus [77]	1.3B	0.16	0.26	0.35	0.28	0.30	0.14	0.23
JanusFlow [53]	1.3B	0.13	0.26	0.28	0.20	0.19	0.11	0.18
MANGO (Ours)	1.3B	0.46	0.52	0.65	0.41	0.48	0.27	0.47
MANGO-4B (Ours)	4.2B	0.50	0.54	0.69	0.46	0.56	0.34	**0.56**

larger models like Janus-Pro and SD 3 by a significant margin across nearly all metrics, and achieves a higher overall score (0.56) compared with T2I-R1 (0.54). Notably, MANGO substantially surpasses peer models of the same scale (1.3B), outperforming Show-o by +0.17 and Janus by +0.24 in terms of overall score. These results further validate the broad applicability of our framework, highlighting our superiority in lightweight model design.

Table 4: Evaluation of image generation on MS-COCO. * indicates that CIDEr is calculated based on 30K randomly sampled data from the validation set, rather than the Karpathy test split set.

Model	Params	FID↓	CLIP Score ↑	CIDEr↑
Text-to-Image Diffusion Model				
SD v1.5 [63]	1B	9.93	30.2	–
SD-XL [58]	3.4B	–	31.0	–
DALL-E 2 [62]	4.5B	10.39	31.4	–
DALL-E 3 [3]	–	–	32.0	–
IF-XL [12]	10.1B	6.66	–	–
Unified Multimodal Model				
DREAMLLM [14]	7B	–	–	115.4
Chameleon [67]	7B	26.74	24.3	120.2
Transfusion [90]	7.3B	6.78	25.5	32.0*
SEED-LLaMA [20]	8B	–	–	123.6
MetaMorph [71]	8B	11.8	–	–
Emu 3 [75]	8B	12.8	31.3	-
LLamaFusion [64]	8B	8.61	24.4	38.4*
LLaVAFusion [64]	8B	8.28	24.7	–
NExT-GPT [78]	13B	10.07	–	124.9
Show-o [81]	1.3B	9.24	–	–
Janus [77]	1.3B	8.5	–	–
MANGO (Ours)	1.3B	7.24	26.8	126.5

MS-COCO Benchmarks. We evaluate MANGO (1.3B) on MS-COCO [43] to further assess its basic image generation capability. As shown in Tab. 4, MANGO achieves an FID [29] of 7.24, outperforming larger models such as DALL-E 2 (10.39) and NExT-GPT (10.07), while obtaining a CLIP score [60] of 26.8, comparable to Transfusion (25.5) despite having fewer parameters. MANGO also attains the highest CIDEr [72] score of 126.5, indicating strong text–image alignment and effective multimodal generation.

4.2 Qualitative Results.

Fig. 1 shows that MANGO overcomes limitations such as concept confusion in complex image generation scenarios. Fig. 3 presents the full process: the

Table 5: Ablation results of Paint-CoT on GenEval.

Setting	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding	Overall↑
T2I Gen. Twice	0.97	0.80	0.58	0.78	0.40	0.47	0.67
Paint-CoT Planning & Acting Only	0.98	0.84	0.66	0.81	0.55	0.58	0.73
Paint-CoT Full Process	0.99	0.86	0.71	0.82	0.60	0.64	0.77

Table 6: Ablation results of Paint-CoT on T2I-CompBench.

Setting	Color	Shape	Texture	Spatial	Non-Spatial	Complex	Overall↑
T2I Gen. Twice	65.15	51.36	64.05	13.42	26.89	32.72	42.26
Paint-CoT Planning&Acting Only	76.77	56.70	71.08	28.36	31.28	39.55	50.62
Paint-CoT Full Process	82.37	59.81	74.21	35.71	34.19	42.78	54.85

planning step enriches details without altering the prompt, the acting step follows the layout accurately, and the reflection and correction steps resolve defects such as structural incompleteness, demonstrating the effectiveness of each stage. Appendix Fig. 4 and Appendix Fig. 7 compare MANGO with various existing models, showing that MANGO maintains strong prompt alignment and achieves superior quality among widely used models such as SD3, SD-XL, Show-o, and JanusFlow. Appendix Fig. 6 provides additional examples under diverse prompts, while Appendix Fig. 5 illustrates further cases for the planning step, such as generation with concise positional constraints.

4.3 Ablation Study

Ablation of Paint-CoT. We assess Paint-CoT 's effectiveness on GenEval and T2I-CompBench as shown in Tab. 5 and Tab. 6. All experiments are based on MANGO (1.3B). To ensure fairness, the baseline model is trained for additional epochs using the pre-training paradigm on the same checkpoint, eliminating biases from extra Paint-CoT training steps. Since the full Paint-CoT process involves two image generations, the baseline performs T2I generation twice at test time and selects the best result. Results show that MANGO already outperforms the baseline with only planning and acting steps, and adding reflection and correction further improves performance, fully validating Paint-CoT 's effectiveness. Reflection-Correction steps yield notable gains even when applied to the outputs of Planning-Acting steps that already deliver corrected results, validating the necessity of Reflection-Correction steps. We provide more ablations to better isolate and validate the effectiveness of Reflection-Correction in Appendix A.

Ablation of FoT. We conducted pre-training experiments under identical settings across three variants: dense design, modality-oriented design, and our MANGO with functionality-oriented FoT. As shown in Tab. 7 on MS-COCO, MANGO outperforms both the dense and modality-oriented variants,

Table 7: Ablation of FoT.

Model	**CIDEr↑**	**FID↓**
Dense	116.2	11.3
Modality-Oriented	121.1	9.56
MANGO (Ours)	126.5	7.24

achieving the best FID (7.24) and CIDEr score (126.5). These results demonstrate that MANGO improves both image understanding and generation through a functionality-disentangled design for visual experts, validating that FoT alleviates conflicts inherent in a single shared visual expert. More detailed experimental description and discussion of functionality conflicts are provided in Appendix E.

Results for Visual Quality. To verify that our approach improves text-to-image alignment without compromising visual quality, we evaluated 300 prompts from GenEval using Aesthetic Score [38] and Human Preference Score v2 (HPS v2) [79]. All experiments are based on MANGO (1.3B). Image triplets were generated for T2I Generation, Planning & Acting, and the Full Paint-CoT Process. As shown in Tab. 8, Aesthetic Scores remain stable despite improvements in text-to-image alignment, while HPS v2 win rates demonstrate consistent human preference for Planning & Acting and Paint-CoT outputs over baseline T2I generations, reflecting enhanced alignment, composition, and detail beyond what Aesthetic Scores capture.

Table 8: Results of visual quality on Aesthetic Benchmarks.

Aesthetic Score

Setting	Score
T2I Gen.	5.98
Planning & Acting	6.02
Full Paint-CoT	6.04

HPS v2 Score (Win Rate)

Setting	Score
Planning & Acting vs. T2I	57.8
Full Paint-CoT vs. T2I	62.3

Results of Reflection-Correction cycles. Our reflection-correction steps can be performed multiple times. However, we find that a single reflection–correction cycle already provides a good balance between performance and efficiency. Executing multiple cycles yields only marginal gains while incurring additional inference cost. In Appendix Tab. 12, we present a comparison on GenEval between running one versus two reflection–correction cycles in Paint-CoT. All experiments are based on MANGO (1.3B). The results show that the improvements from a second cycle are relatively limited, with only minor changes across metrics. We also include qualitative examples for the two-cycle setting in Appendix Fig. 8, where the first correction already meets most requirements and the second correction introduces only subtle adjustments.

5 Conclusions

We introduce MANGO, a unified multimodal model that combines the FoT architecture with a Paint-CoT approach to improve image generation in complex compositional scenarios. FoT disentangles branch functionalities, mitigating conflicts in modality-oriented designs. Paint-CoT structures image generation into planning, acting, reflection, and correction, further pushing the boundaries of complex image generation. A multi-task joint training paradigm ensures effective learning for each step. Experiments across multiple benchmarks demonstrate the effectiveness of our method in substantially improving complex image synthesis.

Acknowledgements

This work is funded by National Natural Science Foundation of China (62576305), Ningbo Youth Science and Technology Innovation Talent Project (2025QL052), and the Alibaba Group through Alibaba Innovative Research Program.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science (2023)
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
5. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
6. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
7. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
9. 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)
10. 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)
11. 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. In: Proc. Annu. Conf. Neural Inf. Process. Systems (2023)
12. DeepFloyd: DeepFloyd IF (2023), https://huggingface.co/DeepFloyd/IF-I-XL-v1.0
13. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
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. In: ICLR (2024)

15. Duan, C., Fang, R., Wang, Y., Wang, K., Huang, L., Zeng, X., Li, H., Liu, X.: Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:2505.17022 (2025)
16. 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: ICML (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. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., Liu, X., Li, H.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. ArXiv **abs/2503.10639** (2025), https://api.semanticscholar.org/CorpusID:276961121
19. 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. arXiv preprint arXiv:2306.13394 (2024)
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. 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)
23. Ghosh, D., Hajishirzi, H., Schmidt, L.: GenEval: An object-focused framework for evaluating text-to-image alignment. In: Proc. Annu. Conf. Neural Inf. Process. Systems (2024)
24. Goole: Gemini 3. November 18 2025 (2025)
25. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
26. 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)
27. Guo, Z., Zhang, R., Tong, C., Zhao, Z., Huang, R., Zhang, H., Zhang, M., Liu, J., Zhang, S., Gao, P., et al.: Can we generate images with cot? let's verify and reinforce image generation step by step. arXiv preprint arXiv:2501.13926 (2025)
28. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems **36**, 66923–66939 (2023)
29. 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. In: Proc. Annu. Conf. Neural Inf. Process. Systems (2017)
30. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
31. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025)

32. Jiang, X., Dong, Y., Wang, L., Fang, Z., Shang, Q., Li, G., Jin, Z., Jiao, W.: Self-planning code generation with large language models. ACM Transactions on Software Engineering and Methodology **33**(7), 1–30 (2024)
33. Jin, Y., Xu, K., Chen, L., Liao, C., Tan, J., Huang, Q., Bin, C., Song, C., ZHANG, D., Ou, W., et al.: Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In: ICLR (2024)
34. Jo, K., Yun, J., Choo, J.: Devil is in the detail: Towards injecting fine details of image prompt in image generation via conflict-free guidance and stratified attention. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23595–23603 (2025)
35. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision. pp. 150–168. Springer (2024)
36. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
37. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023)
38. LAION-AI: Aesthetic predictor. https://github.com/LAION-AI/aesthetic-predictor (2023), gitHub repository
39. Laurençon, H., van Strien, D., Bekman, S., Tronchon, L., Saulnier, L., Wang, T., Karamcheti, S., Singh, A., Pistilli, G., Jernite, Y., et al.: Introducing IDEFICS: An open reproduction of state-of-the-art visual language model, 2023 (2023), https://huggingface.co/blog/idefics
40. Liang, W., Yu, L., Luo, L., Iyer, S., Dong, N., Zhou, C., Ghosh, G., Lewis, M., Yih, W.t., Zettlemoyer, L., et al.: Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. arXiv preprint arXiv:2411.04996 (2024)
41. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19401–19411 (2024)
42. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: The Twelfth International Conference on Learning Representations (2023)
43. 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: ECCV (2014)
44. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with ringattention. arXiv preprint arXiv:2402.08268 (2024)
45. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
46. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
47. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proc. Annu. Conf. Neural Inf. Process. Systems (2024)
48. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)

49. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: MMBench: Is your multi-modal model an all-around player? In: ECCV (2024)
50. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
51. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: CVPR. pp. 26439–26455 (2024)
52. Luo, G., Yang, X., Dou, W., Wang, Z., Liu, J., Dai, J., Qiao, Y., Zhu, X.: Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24960–24971 (2025)
53. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Zhao, L., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. arXiv preprint arXiv:2411.07975 (2024)
54. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431 (2024)
55. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., et al.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)
56. OpenAI: GPT-4 technical report. arXiv:2303.08774 (2023)
57. OpenAI: Introducing gpt-5. August 7 2025 (2025)
58. 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. In: ICLR (2024)
59. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025)
60. 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: ICML (2021)
61. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models (2020)
62. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
63. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
64. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Llamafusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188 (2024)
65. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: CVPR. pp. 14398–14409 (2024)
66. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative pretraining in multimodality. In: ICLR (2024)
67. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)

68. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
69. Team, G.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
70. Team, O.: Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. July 4 2024 (2024)
71. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164 (2024)
72. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
73. Wang, C., Lu, G., Yang, J., Huang, R., Han, J., Hou, L., Zhang, W., Xu, H.: Illume: Illuminating your llms to see, draw, and self-enhance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21612–21622 (2025)
74. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
75. 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)
76. 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)
77. 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. arXiv preprint arXiv:2410.13848 (2024)
78. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: NExT-GPT: Any-to-any multimodal LLM. In: ICML (2024)
79. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
80. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: VILA-U: A unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429 (2024)
81. 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. arXiv preprint arXiv:2408.12528 (2024)
82. 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. arXiv preprint arXiv:2408.12528 (2024)
83. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models, 2025. URL https://arxiv. org/abs/2506.15564 (2025)
84. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440 (2024)
85. Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., Luo, J.: Tap: Text-aware pre-training for text-vqa and text-caption. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8751–8761 (2021)

86. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
87. 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)
88. Zhang, Z., Zeng, Z., Lin, Y., Xiao, C., Wang, X., Han, X., Liu, Z., Xie, R., Sun, M., Zhou, J.: Emergent modularity in pre-trained transformers. arXiv preprint arXiv:2305.18390 (2023)
89. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
90. 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. arXiv preprint arXiv:2408.11039 (2024)
91. 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. arXiv preprint arXiv:2408.11039 (2024)
92. Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q.V., et al.: Least-to-most prompting enables complex reasoning in large language models. In: The Eleventh International Conference on Learning Representations (2023)
93. Zhu, Y., Zhu, M., Liu, N., Ou, Z., Mou, X., Tang, J.: LLaVA-Phi: Efficient multi-modal assistant with small language model. arXiv preprint arXiv:2401.02330 (2024)