ThinkGen: Generalized Thinking for Visual Generation

Siyu Jiao[1*] Yiheng Lin[1*] Yujie Zhong[2†] Qi She[2] Wei Zhou[2] Xiaohan Lan[2]
Zilong Huang[2] Fei Yu[2] Yingchen Yu[2] Yunqing Zhao[2] Yao Zhao[1] Yunchao Wei[1†]

[1] Beijing Jiaotong University [2] Bytedance

Home: https://github.com/jiaosiyuu/ThinkGen
HF: https://huggingface.co/JSYuuu/ThinkGen

Abstract

Recent progress in Multimodal Large Language Models (MLLMs) demonstrates that Chain-of-Thought (CoT) reasoning enables systematic solutions to complex understanding tasks. However, its extension to generation tasks remains nascent and limited by scenario-specific mechanisms that hinder generalization and adaptation. In this work, we present ThinkGen, the first think-driven visual generation framework that explicitly leverages MLLM's CoT reasoning in various generation scenarios. ThinkGen employs a decoupled architecture comprising a pretrained MLLM and a Diffusion Transformer (DiT), wherein the MLLM generates tailored instructions based on user intent, and DiT produces high-quality images guided by these instructions. We further propose a separable GRPO-based training paradigm (SepGRPO), alternating reinforcement learning between the MLLM and DiT modules. This flexible design enables joint training across diverse datasets, facilitating effective CoT reasoning for a wide range of generative scenarios. Extensive experiments demonstrate that ThinkGen achieves robust, state-of-the-art performance across multiple generation benchmarks.

1. Introduction

Recent advances in Large Language Models (LLMs) [41, 58, 63] and Multimodal Large Language Models (MLLMs) [16, 36, 45] have demonstrated the effectiveness of Chain-of-Thought (CoT) reasoning, where models generate explicit intermediate steps to systematically solve complex tasks. CoT reasoning has significantly improved performance in areas such as mathematics, coding, and vision-language understanding. Building on these successes, researchers are now increasingly exploring how CoT reasoning can be leveraged to enhance generation tasks.

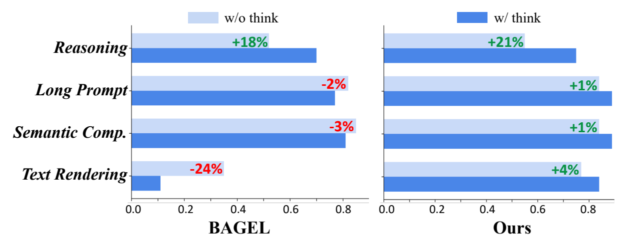

Figure 1. Comparison between BAGEL [10] and our ThinkGen. ThinkGen achieves superior performance when adopting CoT reasoning (w/ think) across a wide range of generation scenarios.

Currently, CoT for generation remains at a preliminary stage. Pioneering work [17] conceptualizes the progressive generation of image tokens as a form of CoT like textual token generation, and focuses on optimizing this process. Recent studies [10, 21, 39, 50] advance the field by refining generation instructions [10] or decomposing the generation process into distinct steps [21, 39, 50], thereby improving image quality in specific tasks. Despite these advances, current methods are constrained by a significant challenge: their CoT mechanisms are typically tailored to a single scenario, *e.g.*, reasoning generation, and may degrade performance when applied to broader tasks (Fig. 1 left). As a result, these approaches typically require manual intervention to activate CoT reasoning for different generation tasks, preventing their flexibility across diverse scenarios.

We attribute the aforementioned challenges to the fact that current frameworks often lack advanced reasoning capabilities. In this work, we introduce ThinkGen, the first think-driven visual generation framework that explicitly leverages a Multimodal Large Language Model (MLLM) with <think> formatting, endowing the system with robust reasoning abilities. This is followed by a dedicated Diffusion Transformer (DiT) for high-quality image synthesis. A key challenge lies in filtering out redundant information

*Equal Contribution. † Corresponding authors.

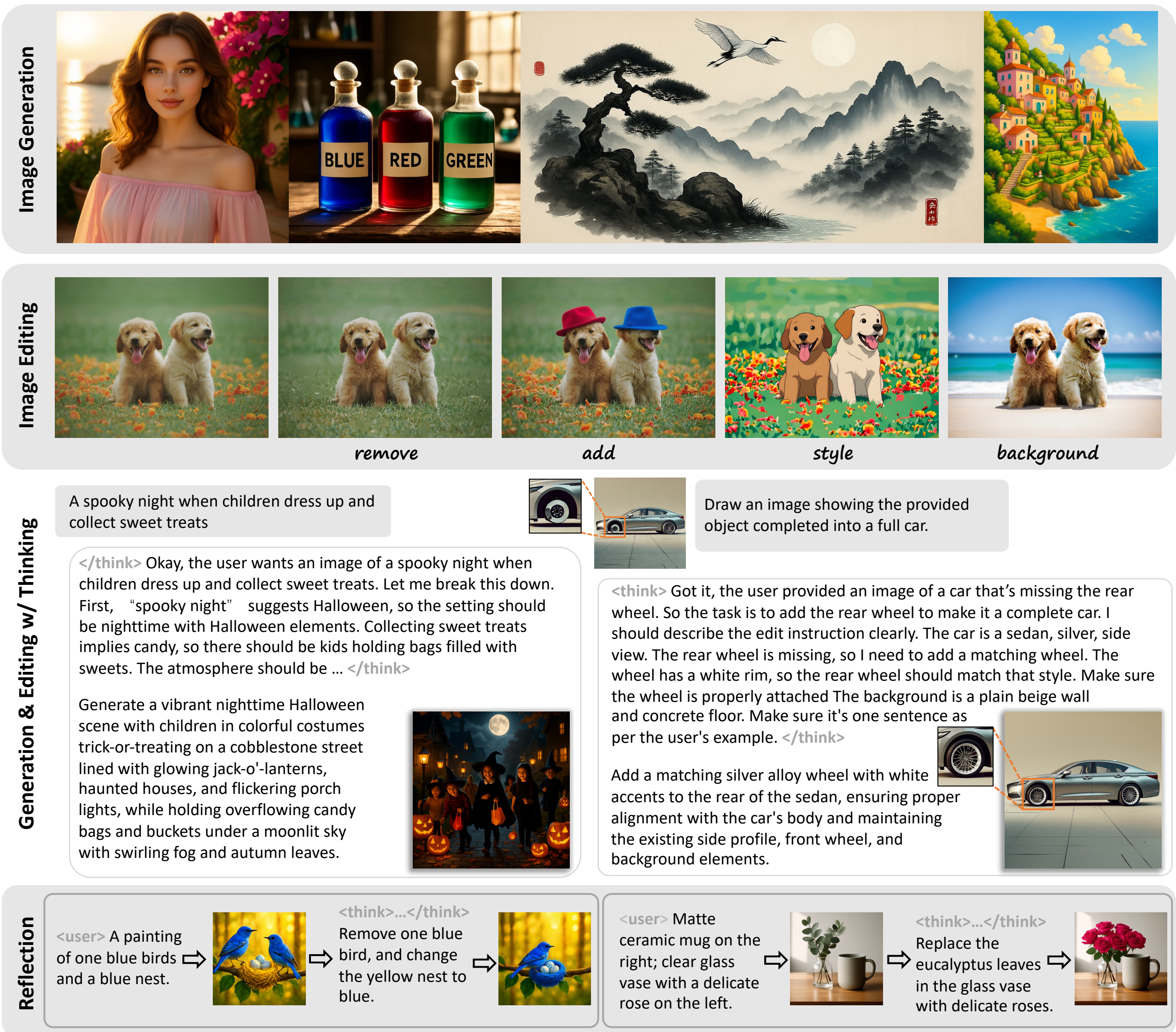

Figure 2. ThinkGen enables think-driven generation across a wide range of scenarios, including text-to-image generation, text rendering, image editing, reasoning generation, reasoning editing, and reflection.

from the chain-of-thought (CoT) reasoning process to make it suitable for guiding the DiT. To this end, we introduce the Visual Generation Instruction refinement (VGI-refine) module, which extracts concise instruction information from the MLLM's reasoning chain and concatenates it with learnable Prepadding States. This enables adaptive adjustment of the MLLM's representation distribution, ensuring better alignment with the requirements of the DiT.

Our training paradigm combines supervised learning with reinforcement learning. In the supervised learning stage, we develop a data-template to generate pseudo-CoT annotations from image-text pairs, addressing the lack of explicit `<think>` labels in most existing generation datasets, enabling the DiT to be optimized in a reasoning-driven manner. In the RL stage, we introduce a separable GRPO-based training paradigm (SepGRPO), where GRPO is applied separately to the MLLM and DiT modules. SepGRPO first freezes DiT while optimizing the MLLM, and then reverses the process by training DiT with the MLLM held fixed. To enhance generalization, we incorporate multi-scenario training data, jointly training the entire model across diverse datasets to achieve robust CoT reasoning in a wide range of generation tasks. By performing the separable design, several advantages are provided: 1) Flexible Reward Design: Distinct rewards can be tailored for each module, enabling more targeted and effective optimization. 2) Reduced Learning Complexity: The MLLM focuses on providing instructions that are well-aligned with DiT's preferences, while DiT specializes in producing high-quality images based on these tailored instructions. 3) Lower Training Cost: The separate design significantly reduces GPU memory usage during training, greatly enhancing computational efficiency.

We evaluate ThinkGen across various generation scenarios. Extensive experiments demonstrate that ThinkGen achieves robust performance on diverse generation benchmarks, *e.g.*, GenEval (0.89), CVTG (0.84), and ImgEdit (4.21). Notably, enabling CoT reasoning in ThinkGen yields substantial improvements on reasoning benchmarks: WISE: 0.55→0.76, RISEBench: 3.6→13.0.

2. Related Work

Unified Model for Generation and Understanding. Recently, building unified models for both generation and understanding has attracted significant attention. Leveraging the strong multimodal understanding capabilities of MLLMs across images and text, image generation has seen further improvements. One line of work [9, 49, 52] adopts VQGAN-style tokenizers [12] and trains MLLMs to generate discrete visual tokens, producing images via next-token prediction in an autoregressive manner. [8, 26, 37, 53, 54] integrates MLLMs with text-to-image diffusion models [23, 38]. The powerful MLLMs are used to extract semantic features, which are then fed as conditions to a diffusion model for image generation. [54] uses the last hidden states as conditional features for generation, while [37] introduces learnable queries to extract informative features for conditioning. However, these methods primarily treat the MLLM as a feature extractor, without fully leveraging its CoT reasoning capabilities. Additionally, some works [4, 10, 33, 60] fuse autoregressive and diffusion modeling. This paradigm autoregressively generates text tokens while producing image tokens via a multi-step diffusion process, combining the strengths of both approaches.

Reinforcement Learning. Recently, reinforcement learning (RL) has been used to enhance MLLMs and diffusion-based generative models. Online RL [16, 20, 40] for MLLMs has been effective at improving MLLMs reasoning capabilities and aligning outputs with human preferences. In particular, [16] shows that rule-based reward functions can elicit human-like, complex chain-of-thought reasoning, while also being memory-efficient by removing the need for a separate value model. [68] utilizes a two-stage training pipeline to enable textual planning in autoregressive generators. Similarly, [21] introduces BiCoT-GRPO to collaboratively optimize high-level semantic planning and low-level token generation. A number of works [28, 31, 62] also apply GRPO to flow-matching models [13, 23] with task-specific rewards. This yields a stable approach for aligning visual outputs with human preferences, improving aesthetics, text rendering, and image–prompt consistency.

3. Model Architecture

We introduce ThinkGen, a think-driven unified model designed for various visual generation tasks, with its architecture shown in Fig. 3. Our model utilizes decoupled MLLM and DiT modules, dedicated to understanding and generation, respectively. This design ensures optimal performance for each component while maintaining both scalability and modularity within the system. For generation tasks, the MLLM receives an image caption or reference image(s) along with editing instructions as input, and outputs rewritten generation instructions tailored to the preferences of DiT. The DiT module then uses these refined instructions to generate high-quality images.

3.1. Multimodal Large Language Model

As shown in Fig. 3, ThinkGen leverages an MLLM to process both visual and textual inputs, employing autoregressive generation for CoT reasoning. The MLLM is initialized with Qwen3-VL-8B-Think [45]. For image generation tasks, we design a specialized system prompt (`[SYS]`) to encourage the MLLM to understand user intent and provide appropriate rewrite instructions. We then extract the final two layers of hidden states generated after the </think> token as conditional inputs for DiT. Empirical results indicate that using the last two layers of hidden states significantly benefits visual generation.

3.2. Diffusion Transformer

ThinkGen employs a standard DiT architecture [26, 54] initialized with OmniGen2-DiT-4B [54], where the output from the MLLM is used as conditional textual input for generation. In image edit task, additional reference image(s) are processed by a VAE [47] and incorporated as conditional visual inputs. The visual and textual inputs are concatenated with the noisy latent features, enabling joint attention across modalities. We employ a simple linear layer as a connector to align features from multiple conditional inputs. We experimentally find that this straightforward linear projection outperforms MLP-based or more complex transformer-based connectors.

VGI-refine. To address the redundancy in the MLLM's autoregressive chain-of-thought (CoT) outputs [56, 61], we introduce Visual Generation Instruction Refinement (VGI-refine), which consists of two steps. First, instruction tokens following the special token </think> are extracted from the text tokens generated by the MLLM, thereby isolating the essential CoT results for downstream image generation. Second, we concatenate K learnable Prepadding States to the extracted instruction tokens. This concatenation regulates the data distribution of the output hidden states and is especially beneficial for short instructions (*e.g.*, *generate a dog* or *remove the cat*). The resulting refined instruction states are then provided as conditional input to the DiT.

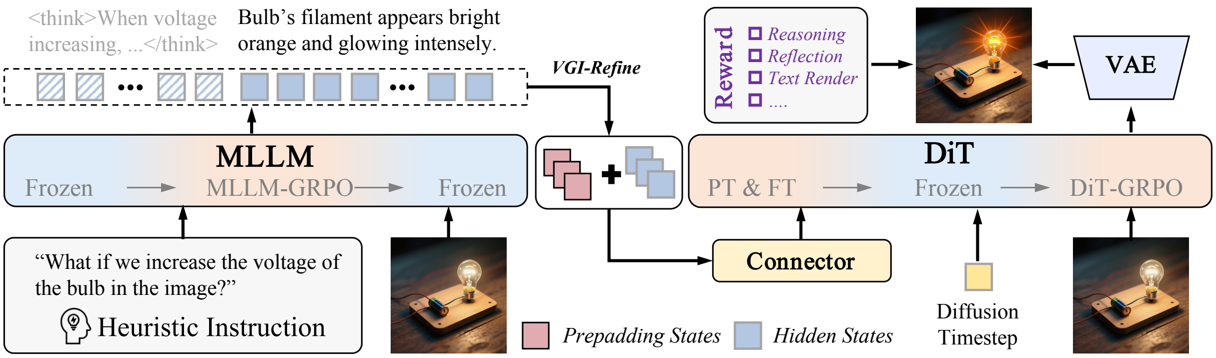

Figure 3. **Overview of ThinkGen.** Within ThinkGen, the MLLM and DiT architectures are decoupled for autoregressive CoT generation and diffusion-based image generation. The MLLM receives text/images as input and outputs generation instructions tailored to the preferences of the DiT. Through a process called visual generation instruction refinement (VGI-Refine), the hidden states corresponding to these instructions are extracted and concatenated with Prepadding States, forming the conditional information for DiT's image generation. For clarity, we omit the text encoder and vision encoder components within MLLM and DiT.

4. Training Recipe

Our ThinkGen training is divided into five distinct stages. Initially, we perform Supervised Pre-training on DiT (Stage 1–3) to ensure high-quality image generation. Subsequently, we introduce a separable MLLM and DiT reinforcement learning approach called SepGRPO (Stage 4–5). Through SepGRPO training, the MLLM learns to generate captions or editing instructions that are optimally aligned with DiT's preferences, while DiT is further refined to produce superior images based on these tailored instructions. The overall training workflow is depicted in Fig. 4.

4.1. Supervised Pre-training

The Supervised Pre-training stages (Stage 1–3) are designed to align the DiT with the MLLM, while simultaneously enhancing image generation quality. We adopt the Rectified Flow [30] training paradigm, which directly regress the velocity field $v_\theta(x_t, t)$ by minimizing the Flow Matching objective [27, 30]:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\, x_0 \sim X_0,\, x_1 \sim X_1} \left[\left\| \mathbf{v} - v_\theta(x_t, t) \right\|^2 \right], \quad (1)$$

here $\mathbf{v} = x_1 - x_0$ denotes the target velocity field.

Input Format. Rewriting each caption or edit instruction during pre-training would be prohibitively expensive. Therefore, in Stage 1–3, we construct pseudo-CoT templates to simulate the MLLM's CoT process. Specifically, we leave the content within `<think> </think>` empty and simply repeat the original caption or edit instruction as the answer. The resulting template is: `[SYS]+[C]+<think> </think>+[C]`, where `[SYS]` denotes the system prompt, and `[C]` denotes the image caption or editing instruction.

Stage1 Alignment. In this stage, we introduce K Learnable prepadding states and align the DiT with the MLLM by training only the linear connector, while keeping the MLLM and DiT frozen. Each image is resized to $\leq 512\times512$px.

Stage2 Pre-training. During this stage, all DiT parameters are trainable. The training corpus comprises 60M image samples, consisting of text-to-image, image edit, text rendering and in-context generation data. Each image is resized to no more than 512×512 pixels.

Stage3 High-quality fine-tuning. In the supervised fine-tuning stage, we construct a 0.7M high-quality subset to enhance DiT's instruction-following capability and image aesthetic. The maximum of training resolution is set to 1024×1024 pixels.

4.2. SepGRPO

We propose SepGRPO, an RL training strategy designed to encourage the MLLM to generate captions/editing instructions that are optimally aligned with DiT's preferences, while enabling DiT to produce higher-quality images based on these instructions. SepGRPO decouples the rollout process for text and vision: first, DiT is fixed while GRPO is applied to the MLLM through joint multi-task training; then, the MLLM is fixed while GRPO is applied to DiT.

Input Format. We design a specialized `[SYS]` during on-policy training to facilitate a cold start, allowing the MLLM to explore text conditions favored by DiT. Specifically, We concatenate the `[SYS]`, the input sample `[C]`, and a special `<think>` token as the input to the MLLM. The resulting template is: `[SYS]+[C]+<think>`.

Stage4 MLLM-GRPO. In this stage, we apply GRPO to the MLLM to encourage the generation of rewritten text that aligns with DiT's preferences. We optimize the MLLM on multiple scenarios to enhance the generalization capability of CoT reasoning. Specifically, we select five representative generation scenarios: semantic composition, reasoning gen-

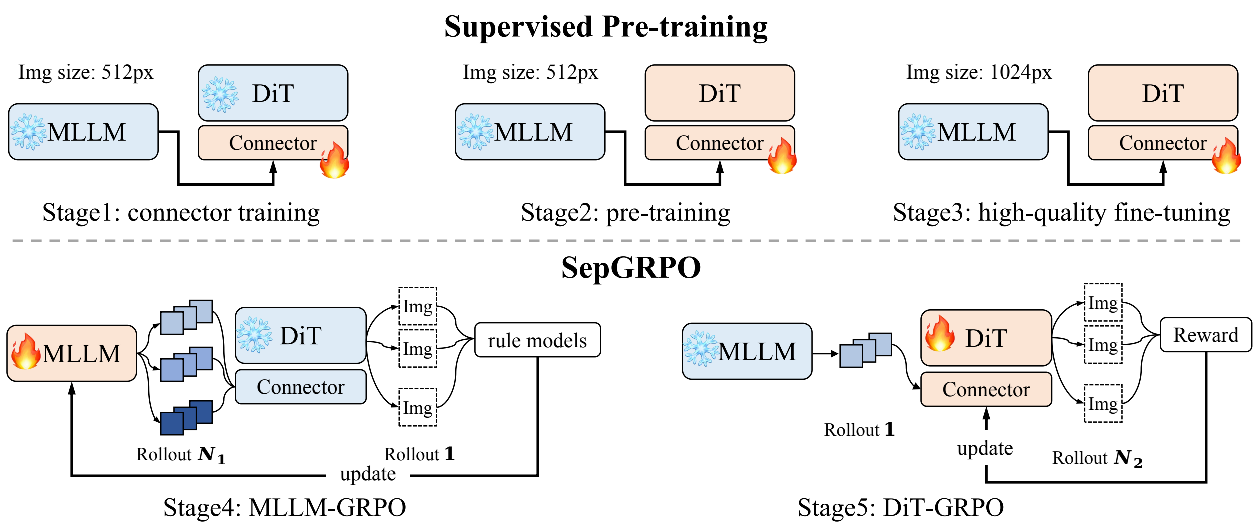

Figure 4. The training recipe of ThinkGen consists of three supervised pre-training stages: Connector training (stage 1), Pre-training (stage 2), and High-quality fine-tuning (stage 3), as well as two SepGRPO stages: MLLM-GRPO (stage 4) and DiT-GRPO (stage 5).

eration, text rendering, image editing, and reflection. For each scenario, we collect and curate dedicated datasets and design corresponding rule models to guide the optimization. The details of the datasets and rule models for each scenario are summarized in Table 1.

Scenario	Dataset	Rule Model
Semantic composition	5K semantic prompts	GenEval [15]
Reasoning generation	10K reasoning prompts	HPSv3 [34]
Text rendering	3K text rendering prompts	Word Acc. [11]
Image editing	3K editing samples	SigLIP2 [46]
Reflection	3K reflection samples	NED

Table 1. Training data and Rule Models in MLLM-GRPO. Notably, all training data and evaluation benchmarks are strictly non-overlapping, ensuring unbiased assessment.

For each input to the MLLM, we perform N_1 rollouts from the policy $\pi_{\theta_{\text{old}}}$ to generate trajectories $\{o_i\}_{i=1}^{N_1}$, which are subsequently used by DiT to produce the corresponding images. Specifically, DiT generates one image for each trajectory. To mitigate the impact of image generation stochasticity, we ensure that all trajectories corresponding to the same input share identical latent noise. The corresponding rule models are then used to calculate a reward $\mathcal{R}_i$ for each trajectory. Subsequently the advantage $\hat{A}_i$ for the i-th trajectory is computed in a group-relative manner:

$$\hat{A}_i = \frac{\mathcal{R}_i - \text{mean}(\{\mathcal{R}_i\}_{i=1}^{N_1})}{\text{std}(\{\mathcal{R}_i\}_{i=1}^{N_1})}. \tag{2}$$

The policy $\pi_{\theta_{\text{old}}}$ is then updated by optimizing the GRPO objective, which is a clipped surrogate function with KL-divergence regularization:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left(\text{MIN} - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right)\right] \tag{3}$$

$$\text{MIN} = \min\left(r_{i,t}(\theta)\hat{A}_i, \text{CLIP}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\right), \tag{4}$$

where $r_{i,j}(\theta)$ denotes the ratio between the probabilities of π_θ and $\pi_{\theta_{\text{old}}}$ for outputting the current token.

In this process, DiT and the rule models jointly serve as reward models. This diverse reward design allows our model to adaptively apply CoT reasoning across a wide range of generation tasks. We provide detailed descriptions of the [SYS], training data distribution, and rule model settings in the appendix.

Stage5 DiT-GRPO: In this stage, we apply FlowGRPO [28] to enhance the instruction-following capability of DiT. We utilize data from the *Simple Scene* and *Text Rendering* scenarios, along with their corresponding reward calculation methods. The training data used in this stage is strictly non-overlapping with that of Stage 4. For each input, the frozen MLLM first performs a single rollout to generate a CoT reasoning trajectory, after which DiT conducts N_2 rollouts to generate N_2 corresponding images. We then compute the advantages as defined in Equation 2 and update the DiT's policy by maximizing the GRPO objective in Equation 3. This process encourages the DiT to favor trajectories that yield higher rewards.

Denoising Reduction: Denoising Reduction [28] (20 steps with 512px) is employed to accelerate the sampling process. This approach enables the efficient collection of low-quality yet informative trajectories during training.

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
GPT-4o [7]	0.81	0.71	0.89	0.83	0.79	0.74	0.80
Gen. Only							
SDXL [38]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD-3.5-large [12]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
FLUX.1-dev [23]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
PixArt-α [5]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Und. and Gen.							
VILA-U [55]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Janus-Pro-7B [9]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3 [49]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Show-o [60]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
MetaQuery-XL [37]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
BLIP3-o-8B [8]	–	–	–	–	–	–	0.62
BAGEL [10]	0.44	0.55	0.68	0.44	0.60	0.39	0.52
BAGEL* [10]	0.76	0.69	0.75	0.65	0.75	0.58	0.70
OmniGen2 [54]	0.42	0.52	0.64	0.43	0.50	0.34	0.47
STAR [25]	0.61	0.67	0.61	0.74	0.69	0.66	0.66
ThinkGen	0.53	0.55	0.71	0.51	0.58	0.40	0.55
ThinkGen*	**0.78**	**0.73**	**0.85**	**0.74**	**0.74**	**0.68**	**0.76**

Table 2. Evaluation of reasoning generation ability on WISE benchmark. * denotes that CoT reasoning is utilized during image generation.

5. Experiments

In this section, we first provide a brief overview of the data composition (Sec. 5.1) and evaluation setup (Sec. 5.2). Next, we evaluate ThinkGen across a variety of visual generation benchmarks (Sec. 5.3). Furthermore, we conduct detailed ablation studies to verify the contribution of each component and training strategy (Sec. 5.4). We also analyze the SepGRPO process (Sec. 5.5) .

5.1. Data composition

For text-to-image generation, our training dataset comprises 54M image-text pairs sourced from publicly available datasets [2, 7, 8, 14, 48]. For image editing tasks, we utilize a diverse set of open-source image editing datasets [7, 22, 32, 42, 51, 64, 70], totaly 5M samples. Furthermore, 1M high-quality proprietary samples are used to further enhance the model's ability to generate visually appealing images and text rendering ability, see the appendix for more details.

5.2. Evaluation setup

Reasoning Generation. We assess reasoning generation capability on WISEBench [35], a world knowledge-informed semantic evaluation benchmark (1000 prompts).
Reasoning Editing. Reasoning editing capability is evaluated on RISEBench [71] (360 pairs). RISEBench evaluates the model's reasoning editing capability across four fundamental types: temporal reasoning, causal reasoning, spatial reasoning, and logical reasoning.
Text-to-image Evaluation. This task evaluates semantic consistency on GenEval [15] (553 prompts) and long-form generation ability on DPG-Bench [18] (1065 prompts), as well as text rendering capability on CVTG [11] (2000 prompts). In the CVTG benchmark, we report the word accuracy of text rendering to assess model performance.
Image Editing Evaluation. We assess image editing capability on ImgEdit [65] (737 pairs), which covers object-level, background, style, and composite manipulations.

Models	Tem.	Cau.	Spa.	Log.	Avg.
Closed-source					
GPT-4o [7]	34.1	32.2	37.0	10.6	28.9
Gemini-2.0 [44]	8.2	15.5	23.0	4.7	13.3
Open-source					
OmniGen [57]	1.2	1.0	0.0	1.2	0.8
EMU2 [43]	1.2	1.1	0.0	0.0	0.5
Step1X-Edit [29]	0.0	2.2	2	3.5	1.9
HiDream-Edit [1]	0.0	0.0	0.0	0.0	0.0
FLUX-Canny [24]	0.0	0.0	0.0	0.0	0.0
BAGEL [10]	3.5	4.4	9.0	5.9	5.8
BAGEL* [10]	5.9	**17.8**	**21.0**	**1.2**	11.9
OmniGen2 [54]	0.0	2.2	7.0	2.3	3.0
ThinkGen	3.5	2.2	7.0	1.1	3.6
ThinkGen*	**16.4**	17.7	16.0	1.1	**13.0**

Table 3. Evaluation of reasoning editing ability on RISEBench.

5.3. Comparison with the state of the art methods

We report results of ThinkGen *w.* and *w.o.* CoT reasoning. When generation *w.o.* CoT reasoning, we simulate the CoT process by adopting the Input Format described in Sec. 4.1.
Reasoning Generation. We conduct experiments on WISEBench to evaluate the reasoning generation capa-

Model	GenEval Counting	GenEval Position	GenEval Overall	DPG Global	DPG Entity	DPG Attribute	DPG Relation	DPG Overall	CVTG Acc.	CVTG NED
Gen. Only										
SDXL [38]	0.39	0.15	0.55	83.27	82.43	80.91	86.76	74.65	-	-
FLUX.1-dev [23]	0.75	0.68	0.82	82.10	89.50	88.70	91.10	84.00	0.49	0.68
PixArt-α [5]	0.44	0.08	0.48	–	–	–	–	–	–	
SD3-Medium [13]	0.72	0.33	0.74	87.90	91.01	88.83	80.70	84.08	0.65	0.84
Sana-1.6B [59]	0.62	0.21	0.66	–	–	–	–	–	–	–
TextCrafter [11]	–	–	–	–	–	–	–	–	0.76	0.90
Und. and Gen.										
Emu3-Gen [49]	0.34	0.17	0.54	85.21	86.68	86.84	90.22	80.60	–	–
ILLUME+ [19]	0.62	0.42	0.72	–	–	–	–	–		
Janus-Pro [9]	0.59	0.79	0.80	86.90	88.90	89.40	89.32	84.19	–	–
MetaQuery-XL [37]	–	–	0.80	–	–	–	–	82.05	–	–
BLIP3-o-8B [8]	–	–	0.84	–	–	–	–	81.60	–	–
BAGEL [10]	0.81	0.64	0.82	88.94	90.37	91.29	90.82	85.07	0.35	0.65
BAGEL* [10]	0.78	0.52	0.79	90.13	90.41	88.73	88.22	83.46	0.11	0.39
OmniGen2 [54]	0.88	0.55	0.80	88.81	88.83	90.18	89.37	83.57	0.52	0.77
ThinkGen	0.81	0.79	0.88	90.32	90.86	91.23	**92.48**	85.14	0.80	0.91
ThinkGen*	**0.84**	**0.80**	**0.89**	**90.87**	**91.36**	**91.77**	91.52	**85.87**	**0.84**	**0.94**

Table 4. Evaluation of text-to-image generation ability on GenEval, DPG and CVTG benchmark.

Model	Add	Adj.	Rep.	Rem.	BG	Sty.	Overall
GPT-4o [36]	4.61	4.33	4.35	3.66	4.57	4.93	4.20
Gen. Only							
MagicBrush [67]	2.84	1.58	1.97	1.58	1.75	2.38	1.90
Instruct-P2P [3]	2.45	1.83	2.01	1.50	1.44	3.55	1.88
AnyEdit [66]	3.18	2.95	2.47	2.23	2.24	2.85	2.45
UltraEdit [70]	3.44	2.81	2.96	1.45	2.83	3.76	2.70
Step1X-Edit [29]	3.88	3.14	3.40	2.41	3.16	4.63	3.06
ICEdit [69]	3.58	3.39	3.15	2.93	3.08	3.84	3.05
Und. and Gen.							
OmniGen [57]	3.47	3.04	2.94	2.43	3.21	4.19	2.96
Janus-4o [6]	3.60	3.25	3.27	2.28	3.32	4.47	3.26
BAGEL [10]	3.56	3.31	3.30	2.62	3.24	4.49	3.20
OmniGen2 [54]	3.57	3.06	3.74	3.20	3.57	**4.81**	3.44
UniWorld [26]	3.82	3.64	3.47	3.24	2.99	4.21	3.26
ThinkGen	4.64	4.12	4.07	**3.95**	**4.31**	4.73	4.14
ThinkGen*	**4.75**	**4.25**	**4.15**	3.49	4.3	4.68	**4.21**

Table 5. Evaluation of image editing on ImgEdit benchmark.

bility. In Tab. 2, we compare ThinkGen with previous well-known generative models and unified generation-understanding models. Our ThinkGen demonstrates a significant advantage over methods based on direct generation. By leveraging CoT reasoning, ThinkGen achieves a substantial improvement of +21% (0.55 $\rightarrow$ 0.76), and establishes a new state-of-the-art performance on WISEBench.

Reasoning Editing. As shown in Tab. 3, activation of the CoT mechanism significantly increases performance on RISEBench (3.6$\rightarrow$13.0) and achieves results competitive with the closed-source model Gemini-2.0.

Text-to-image Generation. In Tab. 4, we present the performance of ThinkGen on the GenEval, DPG-Bench, and CVTG benchmarks. With CoT reasoning, ThinkGen consistently demonstrates improvements across all scenarios, and achieves the best results among many well-known models. These results indicate that ThinkGen possesses strong instruction-following and text-rendering capabilities.

Image Editing. In Tab. 5, we compare the performance of ThinkGen on ImgEdit. Compared with a range of open-source models, ThinkGen shows significantly superior metrics, achieving performance comparable to GPT-4o.

5.4. Ablation Study

Training stage ablations. To understand the effect of each training stage in the ThinkGen, including the Supervised Pre-training and the SepGRPO. We start from a pretrained MLLM and DiT, and gradually apply each training stage (see Tab. 6). We present results on GenEval, WISE, and CVTG, which are used to evaluate instruction-following, reasoning generation, and text-rendering, respectively.

- **Stage1:** Training only the connector yields inferior text-rendering performance (CVTG: 0.28), indicating insufficient fine-grained alignment between MLLM and DiT.
- **Stage2:** Large-scale pre-training results in notable improvements in image quality, with GenEval increasing by 10%, WISE by 9%, and CVTG by 35%.
- **Stage3:** High-quality fine-tuning further enhances image details, resulting in an improvement of +12.0% in CVTG.
- **Stage4:** GRPO applied to the MLLM introduces some representation shift in text conditions, slightly affecting image generation on GenEval (-0.01) and WISE (-0.01). However, incorporating CoT significantly boosts reasoning and generation capabilities (WISE: 0.55 $\rightarrow$ 0.76).
- **Stage5:** DiT-GRPO further enhances image generation

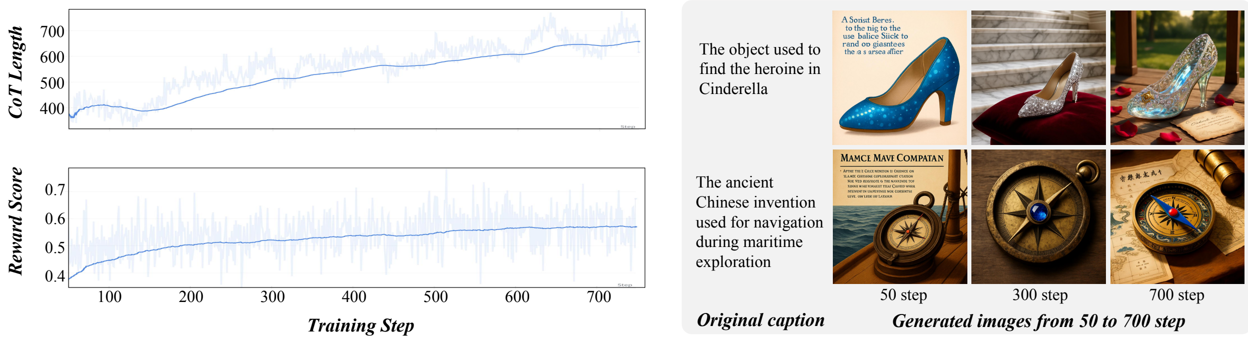

Figure 5. Visualization of the MLLM-GRPO process, the reward score steadily increases.

quality, particularly in fine-grained text rendering tasks. (CVTG: 0.79 → 0.84)

Training stage	GenEval	WISE	CVTG
Stage1 Alignment	0.78	0.46	0.28
Stage2 Pre-training	0.88	0.55	0.63
Stage3 H.Q. Tuning	0.88	0.55	0.75
Stage4 MLLM-GRPO	0.86	0.54	0.75
Stage4 MLLM-GRPO	0.86*	0.76*	0.79*
Stage5 DiT-GRPO	**0.89***	**0.76***	**0.84***

Table 6. Ablation of training stages of ThinkGen. We use the GenEval, WISE, CVTG for analysis. * denotes that cot reasoning is utilized during image generation.

Prepadding States. We compare the results of Stage1 with and without learnable prepadding states in Tab. 7. Prepadding states significantly improve performance on the *short-prompt benchmarks* 0.64→0.78 GenEval, 0.37→0.46 WISEBench, 0.24→0.28 CVTG and 3.46→3.93 ImgEdit, indicating that prepadding states can effectively adjust the representation distribution of MLLM output features and promote alignment between MLLM and DiT.

	Short-Prompt				*Long-Prompt*
	GenEval	WISE	CVTG	ImgEdit	DPG
w.o.	0.64	0.37	0.24	3.46	**80.90**
w.	**0.78**	**0.46**	**0.28**	**3.93**	80.86

Table 7. Ablation of the Prepadding States. We divide the evaluation metrics into *long-prompt* (DPG-Bench) and *short-prompt benchmarks* (GenEval, WISE, CVTG, and ImgEdit) for analysis.

Training strategy. In Tab. 8, we investigate the performance of applying SFT and MLLM-GRPO to the Stage3 model with 10K reasoning data. An interesting phenomenon is observed: directly applying SFT to DiT with reasoning data does not improve performance on reasoning benchmarks, indicating that DiT does not possess the ability to generalize world knowledge to unseen domains. On the other hand, training the MLLM with MLLM-GRPO greatly enhances ThinkGen's reasoning capability (WISE: 0.55 → 0.74). Therefore, the improvement in ThinkGen's reasoning generation capabilities is attributable to the SepGRPO training strategy rather than the reasoning data itself.

	Training data	GenEval	WISE	CVTG
Stage3	-	0.88	0.55	0.75
SFT	10K reasoning data	0.85 −0.03	0.58 +0.03	0.67 −0.08
MLLM-GRPO	10K reasoning data	0.80 −0.08	0.74 +0.19	0.73 −0.02
MLLM-GRPO	24K multitask data	0.86 −0.02	0.76 +0.21	0.79 +0.04

Table 8. Ablation of the Training strategy. GenEval, WISE and CVTG results are used for analysis. Note: both SFT and Text-GRPO are initialized with the model weights from Stage3.

5.5. Analysis of the SepGRPO Process

We visualize the intermediate process of SepGRPO in Fig. 5, including reward scores, CoT length, and generated images. Several key observations emerge: 1) **Increasing CoT Length:** The average CoT length gradually grows, suggesting the model develops more sophisticated reasoning during training. 2) **Unified Reward Growth:** As training progresses, the multi-task reward steadily increases, indicating ThinkGen learns to adaptively think across diverse scenarios. 3) **Image Quality Improvement:** Visualizations at 50, 300, and 700 steps demonstrate a clear trend of improving image generation quality, with generated images exhibiting richer details and higher fidelity.

6. Conclusion

In this work, we introduced ThinkGen, a novel think-driven framework that automatically applies CoT reasoning across diverse generative tasks. Our approach features a decoupled MLLM-DiT architecture trained with SepGRPO, enabling it to formulate a high-quality plan before generation. Extensive experiments demonstrate that ThinkGen achieves significant improvements on reasoning-intensive tasks. Our work represents a key step towards building more intelligent and versatile generative models that seamlessly integrate reasoning and creation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 92470203, U23A20314, 62120106009), and Beijing Natural Science Foundation (No. L242022).

References

[1] Hidream-i1. In *https://github.com/HiDream-ai/HiDream-I1, 2025.* 6

[2] David Beniaguev. Synthetic faces high quality - text 2 image (sfhq-t2i) dataset, 2024. 6

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 7

[4] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, et al. Hunyuanimage 3.0 technical report, 2025. 3

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024. 6, 7

[6] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv:2506.18095*, 2025. 7

[7] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation, 2025. 6

[8] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv:2505.09568*, 2025. 3, 6, 7

[9] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Januspro: Unified multimodal understanding and generation with data and model scaling. *arXiv:2501.17811*, 2025. 3, 6, 7

[10] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv:2505.14683*, 2025. 1, 3, 6, 7

[11] Nikai Du, Zhennan Chen, Shan Gao, Zhizhou Chen, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025. 5, 6, 7

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3, 6

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3, 7

[14] Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025. 6

[15] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 5, 6

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 3

[17] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 1

[18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 6

[19] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv:2504.01934*, 2025. 7

[20] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 3

[21] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 1, 3

[22] Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. NoHumansRequired: Autonomous High-Quality Image Editing Triplet Mining. *arXiv preprint arXiv:2507.14119*, 2025. 6

[23] Black Forest Labs. Flux. *https://github. com/black-forest-labs/flux*, 2023. 3, 6, 7

[24] Black Forest Labs. Flux1-canny. In *https://github.com/black-forest-labs/flux*, 2024. 6

[25] UNIFIED MULTIMODAL LEARNING. Star: Stacked autoregressive scheme for unified multimodal learning. 6

[26] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv:2506.03147*, 2025. 3, 7

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[28] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli

Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 3, 5

[29] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv:2504.17761*, 2025. 6, 7

[30] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4

[31] Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, et al. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025. 3

[32] Jian Ma, Xujie Zhu, Zihao Pan, Qirong Peng, Xu Guo, Chen Chen, and Haonan Lu. X2edit: Revisiting arbitrary-instruction image editing through self-constructed data and task-aware representation learning, 2025. 6

[33] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025. 3

[34] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025. 5

[35] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 6

[36] OpenAI. Introducing 4o image generation, 2025. 1, 7

[37] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv:2504.06256*, 2025. 3, 6, 7

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 3, 6, 7

[39] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025. 1

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[41] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1

[42] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025. 6

[43] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 6

[44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[45] Qwen Team. Qwen3-vl, https://github.com/qwenlm/qwen3-vl. 2025. 1, 3

[46] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[48] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. 2019. 6

[49] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3, 6, 7

[50] Yi Wang, Mushui Liu, Wanggui He, Longxiang Zhang, Ziwei Huang, Guanghao Zhang, Fangxun Shu, Zhong Tao, Dong She, Zhelun Yu, et al. Mint: Multi-modal chain of thought in unified generative models for enhanced image generation. *arXiv preprint arXiv:2503.01298*, 2025. 1

[51] Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025. 6

[52] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025. 3

[53] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3

[54] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv:2506.18871*, 2025. 3, 6, 7

[55] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv:2409.04429*, 2024. 6

[56] Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms, 2025. *URL https://arxiv. org/abs/2502*, 12067. 3

[57] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025. 6, 7

[58] LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025. 1

[59] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv:2410.10629*, 2024. 7

[60] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 3, 6

[61] Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025. 3

[62] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 3

[63] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[64] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 6

[65] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 6

[66] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *CVPR*, 2025. 7

[67] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023. 7

[68] Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl. *arXiv preprint arXiv:2505.24875*, 2025. 3

[69] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv:2504.20690*, 2025. 7

[70] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, 2024. 6, 7

[71] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025. 6