GenAgent: Scaling Text-to-Image Generation via Agentic Multimodal Reasoning

Kaixun Jiang[1], Yuzheng Wang[2]⋆, Junjie Zhou[3], Pandeng Li[2,4], Zhihang Liu[4], Chen-Wei Xie[2], Zhaoyu Chen[1], Yun Zheng[2], and Wenqiang Zhang[1]†

[1] College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2] Tongyi Lab [3] Nanjing University
[4] University of Science and Technology of China

Abstract. We introduce GenAgent, an agentic framework that unifies visual understanding and generation. GenAgent overcomes the limitations of tightly coupled unified models (which suffer from understanding–generation trade-offs and high training data costs) and the structural bottlenecks of existing complex multi-model pipelines through a flexible, decoupled architecture: Understanding is handled by a single multimodal model, while generation is achieved by treating image generation models as invokable tools. Given an image prompt, GenAgent adaptively engages in reasoning, tool invocation, visual judgment, and reflection to iteratively refine outputs until the criteria are met. We employ a two-stage training strategy: first, we cold-start the agent with supervised fine-tuning on high-quality tool-invocation and reflection data to bootstrap multi-turn agent behaviors; second, we perform end-to-end agentic reinforcement learning, combining pointwise rewards (final image quality) and pairwise rewards (reflection accuracy), and utilizing a round-aware trajectory resampling strategy to balance improvements across distinct capabilities. Without modifying the underlying generators, GenAgent significantly boosts the base model (FLUX.1-dev) by 23.6% on GenEval++ and 14.0% on WISE. Beyond substantial performance gains, our framework demonstrates three key emergent properties: (1) cross-tool generalization to generators with varying capabilities, (2) test-time scaling with consistent improvements across interaction rounds, and (3) task-adaptive reasoning. Project page: `https://deep-kaixun.github.io/genagent-page/`

Keywords: Agentic Multimodal Reasoning · Text-to-Image Generation

1 Introduction

Understanding and generation represent the two foundational pillars of multimodal intelligence [8,27]. A growing body of research indicates [4,10,22,23,38,39] that a tight integration of these functions can yield significant gains in complex generative tasks. Current approaches can be broadly categorized into two

⋆ Project leader. † Corresponding author.

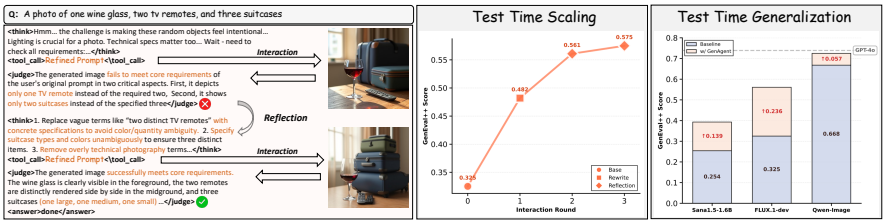

Fig. 1: (Left) Illustration of GenAgent's iterative interaction with image generation tools. (Middle and Right) Performance scaling of our GenAgent across interaction rounds and its zero-shot generalization to different image generation tools.

paradigms: integrated end-to-end models, and modular decoupled systems. The integrated approach involves either incorporating more powerful understanding modules directly into generative models [4, 38] or developing unified architectures [5, 6, 10, 13] capable of processing interleaved image-text inputs to facilitate complex reasoning chains [18, 19, 32]. Nevertheless, such tightly-coupled designs are often constrained by performance trade-offs [26] and demand massive amounts of high-quality interleaved data [10] for training.

Conversely, the decoupled paradigm provides enhanced flexibility by separating semantic understanding from image generation. While early efforts [35, 40] focus on single-turn prompt refinement, recent agent-based methods facilitate multi-turn generation by instantiating distinct cognitive roles. However, these approaches predominantly fall into two categories. The first, including [3, 21, 37], focuses on training-free multi-agent systems that rely on chaining closed-source APIs. These methods manually craft distinct instructions for different functional stages through extensive prompt engineering, leading to rigid reasoning patterns and expensive deployment costs. Alternatively, [48] independently trains distinct functional models (e.g., a corrector and a verifier) to accommodate a Best-of-N sampling strategy. However, its performance heavily relies on drawing a large number of samples, leading to unstable trajectory quality and substantial computational overhead. Overall, existing decoupled frameworks struggle to strike an optimal balance between performance and system complexity.

To address these limitations, we introduce GenAgent, an agentic multimodal framework that achieves functional unification of understanding and generation via a decoupled architecture. Departing from fragmented multi-model pipelines, GenAgent designates a single multimodal model as the central brain to consolidate planning, judging, and reflection. By treating external image generators merely as callable tools, it constructs interleaved multimodal chains-of-thought to autonomously drive multi-turn generation. Specifically, GenAgent formulates plans based on user instructions, invokes tools to generate images, evaluates visual outputs, and iteratively reflects until the criteria are met. Figure 1 (Left) illustrates a representative two-turn interaction.

However, empowering a single model to autonomously navigate this versatile and long-horizon generative process remains highly challenging. To bridge

this gap, we propose a novel two-stage training framework for GenAgent. First, we employ hint-guided distillation to construct high-quality Supervised Fine-Tuning (SFT) data, eliciting accurate tool invocation and reasoning trajectories to cold-start multi-turn agentic behaviors. Subsequently, an end-to-end agentic Reinforcement Learning (RL) phase optimizes the entire long-horizon pipeline—spanning from initial prompt rewriting and intermediate judging, to iterative reflection and final image generation. To mitigate the sparse reward issue inherent in such extended rollouts, we introduce a hybrid reward mechanism combining pointwise outcome rewards for final image quality and pairwise process rewards to incentivize accurate iterative reflections. Finally, to ensure a balanced enhancement of distinct capabilities (e.g. rewriting, reflection), we employ a round-aware resampling strategy. By uniformly sampling across various interaction rounds, this strategy systematically expands exploration space.

We evaluate GenAgent on challenging benchmarks. Without modifying the underlying image generators, GenAgent significantly boosts the base generator (FLUX.1-dev [20]) performance by +23.6% on GenEval++ and +14.0% on WISE. When equipped with stronger generation tools like Qwen-Image [38], our framework approaches the performance of the closed-source GPT-4o [27]. Our analysis reveals three key emergent properties: 1) **Cross-tool generalization**: GenAgent trained through interactions with FLUX.1-dev robustly transfers to other tools with varying capabilities (Figure 1 Right); 2) **Test-time scaling**: performance improves consistently across interaction rounds through multi-turn refinement (Figure 1 Middle); and 3) **Task-adaptive reasoning**: distinct reasoning patterns emerge automatically for different task requirements (Section 4.6). Our contributions are as follows:

- We introduce GenAgent, a flexible agentic multimodal reasoning framework. It consolidates planning, judging, and reflection into a single multimodal model, which interacts with image generators as invokable tools to achieve a deep integration of visual understanding and generation.
- We propose a two-stage training paradigm tailored for long-horizon interaction trajectories. It combines hint-guided SFT for cold-starting with an end-to-end RL framework, utilizing a hybrid reward mechanism and round-aware resampling to ensure the model fully learns and balances distinct capabilities across varying interaction rounds.
- We demonstrate substantial gains on challenging generation tasks, achieving competitive performance with state-of-the-art unified models while uncovering emergent properties including cross-tool generalization, test-time scaling, and adaptive reasoning.

2 Related Work

Understanding for generation. Current approaches to enhancing understanding for controllable generation fall into two paradigms: integrated end-to-end models and modular decoupled systems. The integrated paradigm incorporates stronger understanding modules [4, 34, 38] or builds unified archi-

tectures [6, 10, 13, 39, 42] for complex reasoning. However, these tightly-coupled designs often face performance trade-offs, demand massive interleaved data, and lack flexibility (e.g., further enhancing a specific capability inevitably requires continued training). Conversely, the modular paradigm separates reasoning and generation. While early efforts like RePrompt [40] employed simple prompt optimization, recent agent-based methods expand upon this by constructing complex operational pipelines. They either rely on rigid prompt engineering to chain closed-source APIs [3,21,37], or independently train separate functional modules for expensive Best-of-N (BoN) sampling [48]. GenAgent addresses these limitations by elevating the modular paradigm into an agentic multi-modal model. It adaptively formulates plans based on user instructions, invokes tools, evaluates outputs, and iteratively reflects until criteria are met, substantially expanding the performance ceiling of modular approaches.

Agentic multimodal reasoning. Agentic reasoning has emerged as a powerful paradigm for complex multimodal tasks. In visual understanding, the "thinking with image" concept, pioneered by OpenAI's O3 [28], empowers models with external tools to refine reasoning—evolving from single-function tools like [47] to writing Python code for diverse image manipulations like [17,45]. In contrast, multimodal generation has predominantly relied on unified models [11,18,19,25, 29, 32], which integrate reasoning and generation within a single architecture. However, this tightly-coupled design limits flexibility and scalability: enhancing one capability (e.g., generation quality) requires costly retraining of the entire model. Our work addresses this limitation by pioneering a decoupled agentic framework for generation. An agentic mulitimodal model serves as the reasoning agent while image generators function as callable external tools. This design enables significant advantages: generative performance can be directly improved by upgrading tool components without retraining the core agentic structure, ensuring superior flexibility and scalability.

3 Method

3.1 Overall Pipeline

Overall. Our GenAgent is an agentic framework for multimodal generation. As shown in Figure 2, it comprises two core components: a **multimodal model** π_θ and an **interactive image generation tool**. Within each inference trajectory, GenAgent iteratively invokes the generation tool, orchestrating cycles of **thinking, generation, judgment, and reflection** until the output satisfies user requirements or reaches maximum interaction rounds $n_{\max}$.

Initial Phase. GenAgent first receives user query q and reasons about the intent to produce a refined prompt P_1 in tool-callable format:

$$T_1, P_1 = \pi_\theta(q), \tag{1}$$

where T_1 denotes the reasoning trace. The prompt P_1 is passed to the image generation tool to produce $I_1 = \text{ImageGen}(P_1)$.

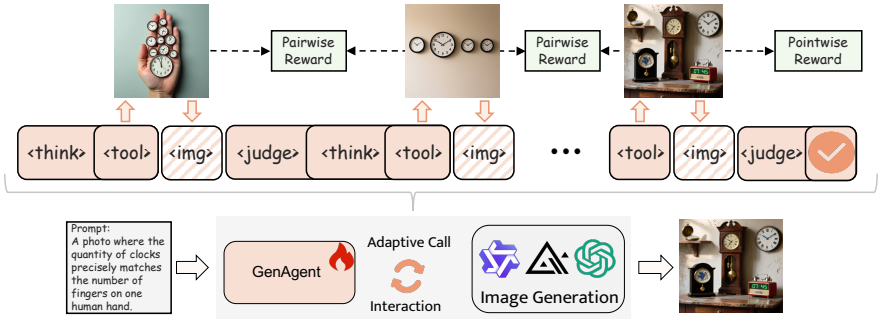

Fig. 2: Overview of GenAgent. Upon receiving user input, GenAgent executes a multimodal reasoning trace that encompasses iterative reasoning and image generation through dynamic interaction with image generation tools. During the rollout of agentic reinforcement learning, we compute pointwise rewards from the final image and pairwise rewards based on consecutive image pairs throughout trace.

Judgment & Reflection. Then, GenAgent evaluates I_1 against user requirements:

$$\pi_\theta(q, T_1, P_1, I_1) = \begin{cases} (J_1, a), & \text{if } I_1 \text{ satisfies } q \\ (J_1, T_2, P_2), & \text{otherwise} \end{cases} \tag{2}$$

where J_1 is the judgment trace and a denotes termination. If the result is satisfactory, the system outputs a. Otherwise, J_1 identifies deficiencies and T_2, P_2 guide the next iteration. This process continues until satisfaction or maximum rounds $n_{\max}$ is reached. A complete multimodal reasoning trajectory (corresponding to the upper part of Figure 2) is:

$$o = \{q, T_1, P_1, I_1, J_1, \ldots, T_n, P_n, I_n, J_n, a\}, \tag{3}$$

where $n \leq n_{\max}$ denotes the actual number of interaction rounds. o captures the complete multimodal chains-of-thought and self-correction capability. The final generated image I_n serves as the output.

3.2 Supervised Fine-Tuning

We first investigate whether multimodal models possess sufficient zero-shot capability for this task. Specifically, we evaluate Qwen2.5-VL-7B [2] and identify three critical limitations through our diagnostic analysis on GenEval++ (Table 1): 1) **Unreliable tool invocation**, indicated by a high Error Rate (the tool invocation failure rate); 2) **Poor refinement**, reflected by the Word Diffs metric (meaning the word count difference between the final prompt and user input is often ≤ 5); and 3) **Ineffective reflection**, evidenced by a negligible Improvement (the performance gain attributed to reflection). These limitations indicate that direct RL would be inefficient: invocation errors prevent effective interaction with generators, weak reflection hinders learning from feedback, and

Table 1: Diagnostic analysis of Base and SFT models on GenEval++ [43].

Stage	Error Rate(%)	Word Diffs ≤ 5 (%)	Improvement(%)
Base	13.36	49.28	0.36
$+SFT$	0.35	0.00	6.07

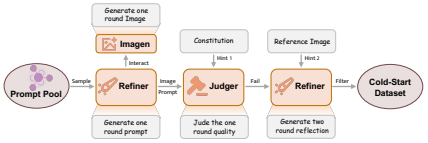

Fig. 3: SFT data construction pipeline.

poor refinement leads to limited RL exploration. Therefore, a cold-start phase is essential.

Figure 3 illustrates the complete trajectory synthesis pipeline for cold-start training data. Data sources, and distributions and more details are provided in the Supplementary Materials.

Prompt Pool. To enable the model to effectively learn tool invocation and reflection capabilities, we carefully curate sufficiently challenging training data. Specifically, we first collect open-source data, then construct synthetic data requiring complex reasoning to match the difficulty of multi-turn tool invocation (detailed pipeline in Supplementary Materials; Figure 2 shows a synthetic case with original prompt 'a photo of five clocks'). We filter these through FLUX.1-dev [20] with 3-pass generation, retaining only samples where all attempts fail.

One-round Trajectory Synthesis. The first round aims to teach the model the format for reasoning and invoking the image generation tool. We employ Qwen3-VL-235B-A22B-Thinking [1] as the teacher model for distillation to obtain the thinking process and the refined prompt P_1. After filtering out format-incorrect samples, we feed all refined prompts to the FLUX.1-dev to produce the corresponding images.

Judging Synthesis. Inspired by [15], we provide explicit rules for the model to judge whether the generated image satisfies the prompt requirements and output the reasoning behind the judgment. This allows the model to implicitly learn these evaluation criteria during training, even though they are not explicitly provided during inference. We retain samples judged as successful, while unsuccessful ones proceed to the two-round reflection.

Reflection Synthesis. We prompt the model to refine the one round P_1 based on the user's original input q and previous evaluation results J_1. To ensure high-quality reflection, we implement two key improvements: First, we upgrade to the more capable Gemini 2.5 Pro [8] as the teacher model. Second, since the seed prompt pool is derived from SFT data, all examples have corresponding reference images. We leverage these reference images as additional guidance for the second-round refinement. We filter out trajectories that explicitly reference the hint image in their prompts, then generate new images using the refined second-round prompts. Through pairwise comparison, we eliminate cases where the second round produces inferior results compared to the first. Finally, we apply pointwise evaluation to assess whether the second-round images satisfy user requirements, retaining only those samples that meet quality standards.

Data Sampling Strategy. Through the above pipeline, we construct two-round training data with tool invocation and reflection. We further filter out samples with logical inconsistencies in the trajectories. Finally, we perform balanced sampling based on data type and whether the second round terminates, ultimately obtaining 32K high-quality cold-start data samples. Notably, we intentionally avoid having all samples end with a to ensure the model is not limited to single-round reflection, thereby preserving exploration space for subsequent reinforcement learning.

Training Strategy. To mitigate hallucination and leverage high-quality trajectories, we mask environment feedback tokens and compute loss only on model responses.Table 1 shows that SFT successfully addresses the base model's limitations: (1) tool invocation errors decrease substantially, (2) refined prompts become longer and more detailed (see Supplementary Material), and (3) reflection quality improves markedly.

3.3 Agentic Reinforcement Learning

After cold-start training, GenAgent has acquired basic tool-calling capabilities. However, SFT relies on fixed round trajectories for imitation learning, making it difficult to generalize to complex scenarios requiring more rounds or to learn when to stop reflection. Therefore, we introduce reinforcement learning to train the model's dynamic decision-making abilities. We adopt GRPO [16]. For each query q, GRPO samples a group of trajectories $\{o_1, \cdots, o_G\}$ from policy π_θ and optimizes via:

$$\mathcal{J}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big\{\min\Big[\mathrm{clip}(\sigma_{i,t}, 1-\epsilon, 1+\epsilon)A_i,\ \sigma_{i,t}A_i\Big]\Big\}, \tag{4}$$

where $\sigma_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ denotes the likelihood ratio between the updated and old policies at step t. ϵ is a hyperparameter. In agentic RL [47], we compute the loss only on the output tokens generated by π_θ, while masking environment observations (i.e., the generated images I in trajectory o). The group-normalized advantage, denoted as A_i, is calculated by the reward r_i:

$$A_i = \frac{r_i - \mathrm{mean}(\{r_i\}_{i=1}^{G})}{\mathrm{std}(\{r_i\}_{i=1}^{G})}. \tag{5}$$

Existing RL methods [16, 44] for reasoning tasks often use rule-based outcome rewards to avoid reward hacking. However, Our task faces two challenges: 1) image quality is difficult to quantify; 2) outcome rewards are too coarse-grained, failing to distinguish the quality of effective reflection trajectories. We propose a hybrid reward mechanism combining pointwise and pairwise rewards.

Pointwise Reward. We employ an advanced MLLM as a Generative Reward Model [24] to verify whether the final generated image satisfies all specified conditions. To ensure evaluation reliability, we implement three key strategies: 1) We prompt the MLLM to first generate step-by-step reasoning before making the

final judgment. 2) We provide detailed evaluation criteria along with condition-specific hints for each sample to guide accurate assessment. 3) We enforce strict rewards where the agent receives a reward only when all conditions are satisfied; otherwise, the reward is 0.

Pairwise Reward. To encourage effective reflection while preventing reward hacking [12], we introduce process rewards based on a key insight: genuine reflection should yield consistent quality improvement across the entire trajectory. We grant rewards only when the MLLM judges that all subsequent images are superior to their predecessors:

$$r_{\text{pair}} = \begin{cases} 0.3 & \text{if } M_{\text{judge}}(I_k, I_{k+1}) = I_{k+1}, \forall k \in [1, n-1], \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $M_{\text{judge}}(I_k, I_{k+1})$ returns the better image between consecutive pairs. To avoid positional bias, the positions are randomly shuffled. The final reward is:

$$r = r_{\text{point}} + r_{\text{format}} + \lambda \cdot r_{\text{pair}}, \tag{7}$$

where $r_{\text{point}} \in \{0, 0.7\}$, $r_{\text{format}} \in \{0, -0.2\}$, and λ is set to 0.5 if $r_{\text{point}} = 0$, otherwise $\lambda = 1$.

Resampling on Interaction Rounds. In standard GRPO, unconstrained rollouts often bias the policy toward specific trajectory lengths, such as premature termination for immediate rewards or meaningless looping. In our framework, however, trajectory lengths correlate with distinct cognitive skills: single-turn interactions cultivate instruction comprehension and prompt rewriting, while multi-turn interactions foster introspective judgment and iterative reflection. This imbalance hinders holistic capability development.

To ensure balanced skill acquisition, we introduce Round-Aware Trajectory Resampling. Inspired by [30], we first oversample G' trajectories ($G' > G$) per query during the rollout phase. Instead of standard unconstrained selection, we construct the final optimization batch of size G by uniformly sampling across subsets categorized by their number of tool invocations. This mechanism systematically expands the exploration space and prevents the policy from collapsing into single-mode reasoning.

4 Experiments

4.1 Implementation Details

We initialize our approach using Qwen2.5-VL-7B [2] as the base model. The supervised fine-tuning (SFT) is performed via LLaMAFactory [46] for 3 epochs utilizing the AdamW optimizer, employing a batch size of 256 and a learning rate of 1×10^{-5}.In the subsequent reinforcement learning (RL) phase, we implement an enhanced GRPO utilizing a batch size of 240 for one epoch. For each prompt, we generate $G' = 12$ rollouts, which are then downsampled to $G = 8$ samples. The KL divergence coefficient is set to 0.0 [44], the maximum response length is

Table 2: Performance Comparison on GenEval++. **bold** indicates the best result.

Type	Method	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
Diffusion	FLUX.1-dev	0.400	0.600	0.250	0.250	0.075	0.400	0.300	0.325
	Qwen-Image	**0.875**	0.725	**0.725**	0.600	0.475	**0.725**	0.550	0.668
Unified	GPT4o	0.900	0.675	0.725	0.625	0.600	0.800	0.850	0.739
	Janus Pro 7B	0.450	0.300	0.125	0.300	0.075	0.350	0.125	0.246
	T2I-R1	0.675	0.325	0.200	0.350	0.075	0.250	0.300	0.311
	Bagel	0.325	0.600	0.250	0.325	0.250	0.475	0.375	0.371
Decoupled	PromptEnhancer	0.500	0.625	0.225	0.375	0.125	0.450	0.375	0.382
	ReflectionFLow	0.400	0.625	0.275	0.275	0.200	0.425	0.325	0.361
	T2I-Copilot	0.500	0.650	0.550	0.325	0.500	0.475	0.475	0.496
	Ours *w/o* tool	0.650	0.575	0.350	0.450	0.350	0.600	0.400	0.482
GenAgent (Ours)	Base	0.300	0.600	0.325	0.300	0.325	0.300	0.350	0.357
	+SFT	0.500	0.550	0.550	0.525	0.250	0.600	0.575	0.507
	+RL	0.750	0.675	0.375	0.525	0.450	0.625	0.525	0.561
	w/ Qwen-Image	0.775	**0.775**	0.650	**0.800**	**0.600**	**0.725**	**0.750**	**0.725**

capped at 16,384 tokens, and a learning rate of 1×10^{-6} is adopted. We leverage the 8-step distilled version of FLUX.1-dev [9] as our interactive image generation tool, with a maximum number of interactions $n_{\max} = 3$; all subsequent references to FLUX.1-dev correspond to this specific version. Qwen3-VL-30B-A3B [1] is employed as the generative reward model.

We implement our RL training using `verl` [31]. To manage computationally intensive multimodal workloads (encompassing image synthesis, policy rollouts, and reward calculations), we deploy the image generation tool and reward model as independent interactive servers. Our training setup utilizes 40 NVIDIA A100 GPUs—allocated as 24, 8, and 8 for the policy, image generation, and reward servers, respectively—whereas our inference requires only 2 GPUs (one for the policy, one for generation).

4.2 Benchmarks

We select three challenging benchmarks to evaluate across multiple dimensions: **1) Instruction-following:** GenEval++ [43], a more accurate and challenging benchmark for evaluating instruction fidelity in image generation, featuring richer semantics and more diverse compositions than the original GenEval [14]. **2) Knowledge-grounded reasoning:** WISE [26], which comprehensively assesses complex semantic understanding and world-knowledge integration in text-to-image generation. **3) Creative generation:** Imagine [43], designed to evaluate surreal and imaginative image generation capabilities by testing model's ability to augment common objects with fantastical elements while preserving their core identity—often requiring outputs that contradict factual reality.

4.3 Performance Comparison

We selected representative baselines spanning different paradigms: 1) Diffusion models: Qwen-Image [38] and FLUX.1-dev [20]; 2) Unified architectures: the autoregressive Janus-Pro-7B [6], reasoning-focused T2I-R1 [19], hybrid-architecture

Table 3: Performance Comparison on WISE. **bold** indicates the best result.

Type	Method	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Diffusion	FLUX.1-dev	0.56	0.57	0.67	0.45	0.54	0.41	0.55
	Qwen-Image	0.62	0.63	0.77	0.57	0.75	0.40	0.62
Unified	GPT4o	0.81	0.71	0.89	0.83	0.79	0.74	0.80
	Janus Pro 7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
	T2I-R1	0.56	0.55	0.63	0.54	0.55	0.30	0.54
	Bagel	0.44	0.55	0.68	0.44	0.60	0.39	0.52
	Bagel *w/* Self-CoT	0.76	**0.69**	0.75	0.65	**0.75**	**0.58**	0.70
Decoupled	PromptEnhancer	0.54	0.60	0.69	0.46	0.62	0.39	0.56
	T2I-Copilot	0.71	0.66	0.77	0.56	0.67	0.48	0.67
	Ours *w/o* tool	0.75	0.66	0.72	0.55	0.60	0.49	0.67
GenAgent (Ours)	Base	0.70	0.63	0.68	0.53	0.61	0.41	0.63
	+SFT	0.68	0.61	0.68	0.60	0.65	0.47	0.64
	+RL	0.75	**0.69**	0.72	0.61	0.65	0.50	0.69
	w/ Qwen-Image	**0.78**	0.67	**0.78**	**0.72**	0.71	0.55	**0.72**

Table 4: Performance Comparison on Imagine. **bold** indicates the best result.

Type	Method	Attribute Shift	Spatiotemporal	Hybridization	Multi-Object	Overall
Diffusion	FLUX.1-dev	5.298	6.350	7.053	5.973	6.072
	Qwen-Image	6.771	7.193	8.130	7.500	7.329
Unified	GPT4o	8.540	9.180	8.570	7.980	8.560
	Janus Pro 7B	5.300	7.280	6.730	6.040	6.220
	T2I-R1	5.850	7.700	7.360	6.680	6.780
	Bagel	5.370	6.930	6.500	6.410	6.200
Decoupled	PromptEnhancer	5.489	6.213	7.327	6.493	6.281
	T2I-Copilot	6.327	6.997	7.243	6.600	6.740
	Ours *w/o* tool	5.829	6.827	7.130	6.137	6.408
GenAgent (Ours)	Base	5.467	6.590	7.287	6.083	6.258
	+SFT	6.282	7.033	7.480	6.527	6.770
	+RL	6.293	7.067	7.543	6.663	6.825
	w/ Qwen-Image	**7.613**	**7.547**	**8.343**	**7.763**	**7.794**

Bagel [10], and closed-source GPT-4o [27]; 3) Decoupled approaches: PromptEnhancer (7B version) [35] and T2I-Copilot [3] and ReflectionFlow [48]. These methods and GenAgent use FLUX.1-dev as the default image generation tool. For fair comparison and efficiency, we report results with $n_{\max} = 2$ for all multi-turn baselines and our method. We report our GenAgent across multiple variants: without tool invocation, degenerating to a single-pass rewrite model (*w/o* tool), with SFT (*+SFT*), with RL (*+RL*), and with a stronger tool (*w/* Qwen-Image).

Diffusion Models. GenAgent substantially enhances diffusion model performance across all benchmarks. When FLUX.1-dev is equipped with our GenAgent, it achieves substantial improvements: +0.14 on WISE (0.69 vs. 0.55), +0.236 on GenEval++ (0.561 vs. 0.325), and +0.753 on ImagineBench (6.825 vs. 6.072). Notably, despite RL training with FLUX, GenAgent generalizes effectively to the stronger Qwen-Image generator, achieving new state-of-the-art results among open-source methods. This demonstrates weak-to-strong scalability at the tool level, analyzed in detail in Section 4.4.

Table 5: Ablation study on training stages and components.

Stage	WISE	GenEval++	Imagine
Base	0.63	0.357	6.258
SFT	0.64	0.507	6.770
RL *w/o* r_{pair}	0.68	0.543	6.814
RL *w/o* resample	0.68	0.546	6.821
RL(Ours)	**0.69**	**0.561**	**6.825**

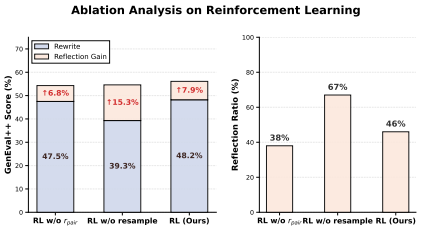

Fig. 4: Ablation analysis on agentic reinforcement learning.

Unified Architectures. On knowledge-grounded reasoning (WISE), GenAgent *w/* FLUX.1-dev achieves 0.69, matching the specialized Bagel *w/* Self-CoT (0.70), while GenAgent *w/* Qwen-Image reaches 0.72, surpassing all open-source unified models and approaching GPT-4o (0.80). For instruction-following (GenEval++), GenAgent with Qwen-Image (0.725) nearly matches GPT-4o (0.739) and significantly outperforms other open-source alternatives. On creative generation (Imagine), GenAgent with Qwen-Image (7.794) substantially narrows the gap with GPT-4o (8.560) while establishing a clear lead over the second-best T2I-R1 (6.780). These results demonstrate that our agentic framework achieves superior generalization and scalability compared to unified architectures, while maintaining competitive performance on complex generative tasks.

Decoupled Methods. GenAgent substantially outperforms existing decoupled approaches. Single-rewrite PromptEnhancer achieves only marginal gains over vanilla FLUX: 0.56 on WISE (+0.01), 0.382 on GenEval++ (+0.057), and 6.281 on Imagine (+0.209). Similarly, workflow-based ReflectionFlow reaches just 0.361 on GenEval++, significantly below GenAgent's 0.561. Furthermore, when compared to training-free multi-agent systems like T2I-Copilot, GenAgent maintains a decisive advantage when fairly evaluated using the same Qwen2.5-VL-7B as the base model. Our approach systematically surpasses T2I-Copilot's complex manual prompt engineering across all metrics, scoring 0.561 vs. 0.496 on GenEval++, 0.69 vs. 0.67 on WISE, and 6.825 vs. 6.740 on Imagine. Notably, even without the agentic framework (i.e., using only a single rewriting, Ours *w/o* tool), our method still outperforms PromptEnhancer across all benchmarks. These results indicate that existing decoupled methods underutilize multimodal model reasoning capabilities. In contrast, GenAgent's end-to-end agentic loop with iterative judgment and reflection unlocks their full potential.

Qualitative Analysis. We provide qualitative analysis, visual comparisons between different methods in Supplementary Materials.

4.4 Ablation Study

Training Stages. As shown in Table 5, the initial Supervised Fine-Tuning (SFT) stage improves instruction-following (GenEval++) and creativity metrics (Imagine) but yields limited gains on the knowledge-intensive WISE benchmark

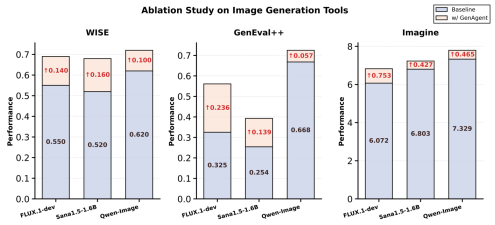

Fig. 5: Ablation study on image generation tools.

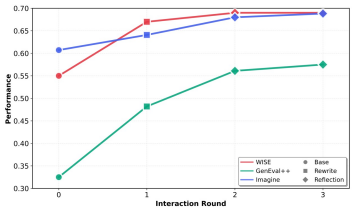

Fig. 6: Ablation on interaction rounds, with normalized scores.

(from 0.63 to 0.64). This is because SFT primarily focuses on learning "how to respond" (e.g., formatting and tool invocation) rather than exploring to activate the model's intrinsic reasoning capabilities. Following this, the agentic Reinforcement Learning (RL) phase effectively elicits these latent reasoning skills, bringing substantial and consistent improvements across all three benchmarks.

Pairwise Reward. We further analyze the impact of our specific RL designs on model behavior. The pairwise reward (r_{pair}) is explicitly designed to supervise and encourage high-quality correction. By comparing the RL *w/o* r_{pair} baseline to ours (RL (Ours)) in Figure 4, we observe that equipping the model with r_{pair} allows it to confidently trigger reflections (the reflection ratio increases from 38% to 46%) while concurrently achieving a higher reflection gain (+6.8% to +7.9%). Moreover, this efficacy is further corroborated by RL *w/o* resample variant (which retains r_{pair}): even as its reflection ratio surges to 67%, the reflection gain scales proportionally to a massive +15.3%. This consistent trend confirms that r_{pair} robustly teaches the agent to translate multi-turn interactions into tangible performance improvements.

Resampling Strategies. While r_{pair} improves reflection quality, naive rollout sampling can inadvertently force the model into a collapsed functional mode. As illustrated in Figure 4, RL *w/o* resample exhibits a severe capability imbalance: driven by the pairwise reward, its reflection ratio heavily spikes to 67% with a massive gain of +15.3%, but at the steep cost of degraded rewriting (rewriting performance crashes to a mere 39.3%), ultimately yielding an inferior overall score. By introducing Round-Aware Resampling, RL (Ours) prevents this pathological capability trade-off. The full model maintains a healthier reflection ratio (46%) and develops robust initial rewriting capabilities (48.2%), while preserving effective correction skills (+7.9%). This unequivocally demonstrates that our resampling strategy harmonizes rewriting and reflection, achieving an optimal balance between interaction efficiency and final performance.

Overall, pairwise reward explicitly promotes highly effective reflection, while the resampling strategy ensures a balanced development of distinct model capabilities. As evidenced in Table 5, the integration of both components produces a strong synergistic effect, enabling our full model to achieve the best performance.

Image Generation Tools. As illustrated in Figure 5, although our GenAgent uses FLUX.1-dev as the image generation tool during reinforcement learning, it

demonstrates exceptional cross-tool generalization and scalability. When replacing the tool with the small Sana1.5-1.6B model [41], GenAgent still achieves significant improvements across all three benchmarks, demonstrating strong backward compatibility. More encouragingly, when employing the more powerful Qwen-Image, GenAgent further enhances its overall performance. This phenomenon indicates that our method can scale with the capability improvements of underlying image generation tools, fully unleashing the potential of advanced generative models.

Interaction Rounds. Figure 6 illustrates the impact of different interaction rounds with the image generation tool on performance. The first interaction demonstrates GenAgent's rewriting capability, with all benchmarks exhibiting significant performance improvements, indicating that initial prompt rewriting effectively comprehends and refines users' original prompts. The second interaction round showcases the effectiveness of the first reflection, with all three tasks showing varying degrees of performance gains, confirming that the reflection mechanism successfully optimizes prompts through introspective analysis. By the third round, marginal gains become negligible: persistent failures stem from either the capability ceiling of the underlying generator on complex semantics and fine-grained attributes, or over-reflection on ambiguous prompts, where the agent oscillates among valid interpretations and degrades earlier outputs. The experimental results indicate that two interaction rounds represent the optimal balance between performance and computational efficiency.

4.5 Diagnostic Analysis

Comparison with Best-of-N (BoN). We compare against a robust BoN baseline, which generates N candidate images using our initial rewritten prompt and employs our training-stage reward model as the verifier to select the optimal output. As shown in Table. 6, GenAgent (Turns=2) significantly outperforms BoN ($N = 2$) by 5.03%. While BoN relies on inefficient stochastic resampling, GenAgent leverages GRPO to distill the expensive *pass*@N search into a transparent *pass*@1 policy. By executing logic-driven prompt updates, GenAgent ensures systematic refinement without requiring external verifiers during inference. Crucially, this experiment also validates our **reward accuracy**. The 2.87% performance boost of BoN ($N = 2$) over BoN ($N = 1$) confirms the reward's precision in distinguishing superior image-text alignments—a consistent gain that would be impossible with a misaligned signal.

Table 6: Comparison with Best-of-N.

Method	Budget	GenEval++ (%)
BoN ($N = 1$)	1 Img	48.20
BoN ($N = 2$)	2 Imgs	51.07
GenAgent	2 Imgs	**56.10**

Training Dynamics and Convergence. We analyze the model's behavioral dynamics and convergence during RL training in Supplementary Material.

Inference Time Cost. Detailed analyses regarding the computational overhead of our method during inference are provided in the Supplementary Material.

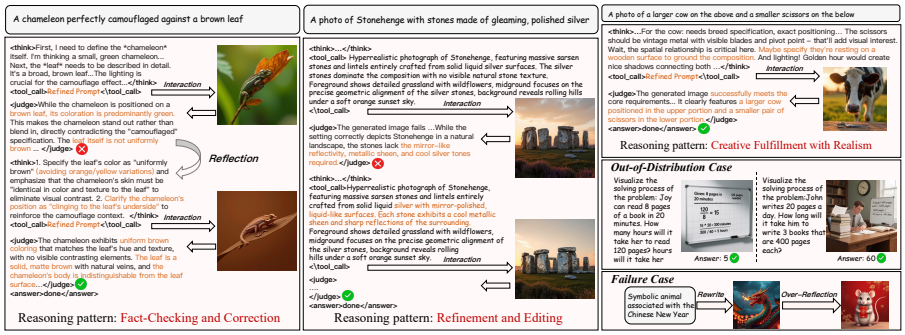

Fig. 7: Case Studies of GenAgent. The figure illustrates three distinct reasoning patterns of GenAgent, its out-of-distribution generalization on a visual math task, and a representative failure case. We provide complete cases in the Supplementary Material.

4.6 Case Study

Key Reasoning Patterns. We highlight three key reasoning patterns in GenAgent's iterative process:
1) Fact-Checking and Correction: In the chameleon example (Figure 7, left), the agent detects a logical conflict—a green chameleon on brown leaves contradicts "perfectly camouflaged"—and revises the instruction to align color and texture, demonstrating self-correction capability.
2) Refinement and Editing: For near-correct outputs like Stonehenge (Figure 7, mid), GenAgent performs targeted edits rather than full regeneration, adding specific descriptors (e.g., "mirror-polished") to efficiently capture missing details like stone gloss.
3) Creative Fulfillment with Realism: The cow-and-scissors case (Figure 7, right-top) shows GenAgent balancing creativity with plausibility by constructing a convincing scene through composition, lighting, and perspective.
Out-of-Distribution Cases. Beyond these patterns, we test GenAgent on out-of-distribution tasks. Inspired by [33], we design a mathematical visualization task using GSM8K [7] problems (Figure 7, right-mid): GenAgent first reasons through the problem, then converts solution steps into detailed visual prompts, demonstrating strong potential for multimodal deep research.
Failure Case. We also identify a failure mode: Over-Reflection, where ambiguous prompts cause the agent to cycle through valid interpretations.

5 Conclusion

In this work, we demonstrate that the functional unification of visual understanding and generation can be elegantly achieved through a decoupled architecture. We introduce GenAgent, an agentic multimodal framework where a single multimodal model serves as the central reasoning brain, treating off-the-shelf image generators merely as callable tools. To effectively optimize this system,

we propose a two-stage training pipeline: hint-guided SFT to cold-start agentic behaviors, followed by agentic reinforcement learning. By leveraging a hybrid reward mechanism and round-aware trajectory resampling, our framework successfully navigates sparse rewards to enable iterative self-refinement. Extensive experiments demonstrate that GenAgent substantially boosts the performance of traditional diffusion models and achieves results comparable to state-of-the-art unified models. Crucially, it maintains greater flexibility and circumvents the requirement of massive interleaved data alignment. Furthermore, our analysis reveals compelling emergent properties, including task-adaptive reasoning, test-time scaling, and zero-shot cross-tool generalization. We hope our work offers an alternative perspective for advancing unified understanding and generation.

Limitation & Future Work. Despite GenAgent's strong performance, two main limitations remain: **1) Reward robustness.** Our reward models are susceptible to reward hacking compared to rule-based alternatives. A promising direction is to decompose holistic image-quality scores into fine-grained, verifiable checklists (e.g., per-attribute satisfaction), so that each reward signal is individually grounded and harder to exploit. Combining such checklist rewards with more robust generative reward models [24, 36] could yield precise, hack-resistant supervision. **2) Tool diversity.** GenAgent currently relies solely on image generation tools for iterative prompt rewriting. Although editing-like behaviors emerge in some cases, explicitly integrating image editing models as callable tools would allow the agent to perform local refinements without full regeneration, improving both efficiency and output quality.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No.62576109), Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.25511104402).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
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. Chen, C.Y., Shi, M., Zhang, G., Shi, H.: T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19396–19405 (2025)
4. 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)
5. Chen, J., Xue, L., Xu, Z., Pan, X., Yang, S., Qin, C., Yan, A., Zhou, H., Chen, Z., Huang, L., et al.: Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857 (2025)
6. 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)

7. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. alimama creative: Flux.1-turbo-alpha. `https://huggingface.co/alimama-creative/FLUX.1-Turbo-Alpha` (2025)
10. 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)
11. 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)
12. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence (2020)
13. Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., et al.: X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. arXiv preprint arXiv:2507.22058 (2025)
14. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
15. Guan, M.Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., et al.: Deliberative alignment: Reasoning enables safer language models. arXiv preprint arXiv:2412.16339 (2024)
16. 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)
17. Hong, J., Zhao, C., Zhu, C., Lu, W., Xu, G., Yu, X.: Deepeyesv2: Toward agentic multimodal model. arXiv preprint arXiv:2511.05271 (2025)
18. Hong, J., Zhang, Y., Wang, G., Liu, Y., Wen, J.R., Yan, R.: Reinforcing multimodal understanding and generation with dual self-rewards. arXiv preprint arXiv:2506.07963 (2025)
19. 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. NeurIPS (2025)
20. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
21. Li, M., Hou, X., Liu, Z., Yang, D., Qian, Z., Chen, J., Wei, J., Jiang, Y., Xu, Q., Zhang, L.: Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13263–13272 (2025)
22. Liao, Z., Jiang, K., Liu, Z., Wei, Y., Yu, J., Li, Q., Yu, H.T., Li, P., Wang, Y., Xing, Z., et al.: Aibench: Evaluating visual-logical consistency in academic illustration generation. arXiv preprint arXiv:2603.28068 (2026)
23. Liu, Z., Bao, X., Li, P., Zhou, J., Liao, Z., He, Y., Jiang, K., Xie, C.W., Zheng, Y., Xie, H.: Showtable: Unlocking creative table visualization with collaborative reflection and refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24405–24416 (2026)

24. Liu, Z., Wang, P., Xu, R., Ma, S., Ruan, C., Li, P., Liu, Y., Wu, Y.: Inference-time scaling for generalist reward modeling. arXiv preprint arXiv:2504.02495 (2025)
25. Lyu, Y., Wong, C.K., Liao, C., Jiang, L., Zheng, X., Lu, Z., Zhang, L., Hu, X.: Understanding-in-generation: Reinforcing generative capability of unified model via infusing understanding into generation. arXiv preprint arXiv:2509.18639 (2025)
26. 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)
27. OpenAI: Gpt-4o. `https://openai.com/index/introducing-4o-image-generation` (2025)
28. OpenAI: Thinking with images. https://openai.com/index/thinking-with-images/ (2025)
29. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., Li, H.: Uni-cot: Towards unified chain-of-thought reasoning across text and vision. arXiv preprint arXiv:2508.05606 (2025)
30. Shang, N., Liu, Y., Zhu, Y., Zhang, L.L., Xu, W., Guan, X., Zhang, B., Dong, B., Zhou, X., Zhang, B., et al.: rstar2-agent: Agentic reasoning technical report. arXiv preprint arXiv:2508.20722 (2025)
31. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)
32. Tong, C., Guo, Z., Zhang, R., Shan, W., Wei, X., Xing, Z., Li, H., Heng, P.A.: Delving into rl for image generation with cot: A study on dpo vs. grpo. arXiv preprint arXiv:2505.17017 (2025)
33. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., et al.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. arXiv preprint arXiv:2511.04570 (2025)
34. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, L., Xing, X., Cheng, Y., Zhao, Z., Li, D., Hang, T., Tao, J., Wang, Q., Li, R., Chen, C., et al.: Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. arXiv preprint arXiv:2509.04545 (2025)
36. Wang, Y., Li, Z., Zang, Y., Zhou, Y., Bu, J., Wang, C., Lu, Q., Jin, C., Wang, J.: Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv preprint arXiv:2508.20751 (2025)
37. Wang, Z., Li, A., Li, Z., Liu, X.: Genartist: Multimodal llm as an agent for unified image generation and editing. Advances in Neural Information Processing Systems **37**, 128374–128395 (2024)
38. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
39. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
40. Wu, M., Wang, L., Zhao, P., Yang, F., Zhang, J., Liu, J., Zhan, Y., Han, W., Sun, H., Ji, J., et al.: Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. arXiv preprint arXiv:2505.17540 (2025)

41. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., Cai, H., Liu, B., Zhou, D., Han, S.: SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer (Mar 2025). https://doi.org/10.48550/arXiv.2501.18427
42. 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)
43. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)
44. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
45. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al.: Thyme: Think beyond images. arXiv preprint arXiv:2508.11630 (2025)
46. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, Bangkok, Thailand (2024), http://arxiv.org/abs/2403.13372
47. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
48. Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., Li, H.: From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. ICCV (2025)