MILR: IMPROVING MULTIMODAL IMAGE GENERATION VIA TEST-TIME LATENT REASONING

Yapeng Mi[1,2] **Yanpeng Zhao**[2 † ✉] **Hengli Li**[2,3] **Chenxi Li**[2,4]
Huimin Wu[2] **Xiaojian Ma**[2] **Song-Chun Zhu**[2,3] **Ying Nian Wu**[5] **Qing Li**[2 ✉]
[1]University of Science and Technology of China
[2]State Key Laboratory of General Artificial Intelligence, BIGAI
[3]Peking University [4]Tsinghua University [5]University of California, Los Angeles
Project: https://spatigen.github.io/milr.io/ **Code:** https://github.com/spatigen/milr

ABSTRACT

Reasoning-augmented machine learning systems have shown improved performance in various domains, including image generation. However, existing reasoning-based methods for image generation either restrict reasoning to a single modality (image or text) or rely on high-quality reasoning data for fine-tuning. To tackle these limitations, we propose MILR, a test-time method that jointly reasons over image and text in a unified latent vector space. Reasoning in MILR is performed by searching through vector representations of discrete image and text tokens. Practically, this is implemented via the policy gradient method, guided by an image quality critic. We instantiate MILR within the unified multimodal understanding and generation (MUG) framework that natively supports language reasoning before image synthesis and thus facilitates cross-modal reasoning. The intermediate model outputs, which are to be optimized, serve as the unified latent space, enabling MILR to operate entirely at test time. We evaluate MILR on GenEval, T2I-CompBench, and WISE; it achieves state-of-the-art results on all benchmarks. Notably, on knowledge-intensive WISE, MILR attains an overall score of 0.63, improving over the baseline by 80%. Our further analysis indicates that joint reasoning in the unified latent space is the key to its strong performance. Moreover, our qualitative studies reveal MILR's nontrivial ability in temporal and cultural reasoning, highlighting the efficacy of our reasoning method.

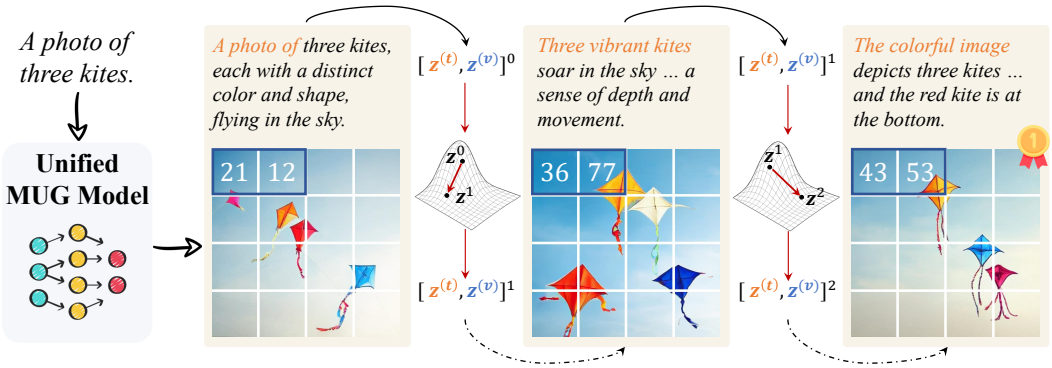


Figure 1: Latent reasoning of MILR. The black *solid* line denotes extracting the output vector representations $\mathbf{z}^k$ of the text tokens $\mathbf{z}^{(t)}$ and image tokens $\mathbf{z}^{(v)}$ to be optimized, and the black *dashed* line denotes decoding from the optimized latent vectors $\mathbf{z}^{k+1}$, where $\mathbf{z} = [\mathbf{z}^{(t)}, \mathbf{z}^{(v)}]$.

Contact: miyapeng78@gmail.com, yannzhao.ed@gmail.com, dylan.liqing@gmail.com.
†: Project Lead. ✉: Corresponding Author.

1 INTRODUCTION

Text-guided image generation is the task of synthesizing an image conditioned on a given text instruction. In recent years, the field has witnessed transformative progress: moving from generative adversarial models (Goodfellow et al., 2014) to autoregressive and diffusion approaches (Sun et al., 2024; Esser et al., 2024; Black Forest Labs, 2024; Chen et al., 2025; Wu et al., 2025a). However, traditional models are limited in generating images in a single-shot fashion and thus are unable to resolve potential defects (Huang et al., 2025a; Niu et al., 2025). Inspired by the success of reasoning-augmented LLMs—such as OpenAI o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025)—that can reflect on and refine their thoughts, and answer accordingly, recent works have attempted to endow image generation models with reasoning ability (Guo et al., 2025; Fang et al., 2025; Zhang et al., 2025b; Wu et al., 2025b; Chern et al., 2025).

Reasoning-augmented image generation models typically perform reasoning in two spaces: language and image spaces. Language reasoning primarily involves refining instructions to make them more comprehensible to the model (Wu et al., 2025b; Li et al., 2025b), and image reasoning iterates on the generation guided by a quality metric (Guo et al., 2025; Zhuo et al., 2025). Early works implement reasoning either in the language or in the image space, lacking a mechanism for synergistic reasoning across the two spaces. To fill the gap, recent works resort to unified multimodal understanding and generation (MUG; Jiang et al. (2025a); Duan et al. (2025); Zhang et al. (2025b)), which natively supports language reasoning before generating images (referred to as multimodal image generation), facilitating cross-modal image-text reasoning. Despite the success, these approaches require carefully curated reasoning data and depend on model fine-tuning, rendering them complex and costly to develop in practice.

To address these limitations, we propose Multimodal Image generation via test-time Latent Reasoning, dubbed MILR. Our core idea is to reason in a unified *latent* vector space that encodes both text and images, starkly different from previous methods that explicitly reason over raw images and text. We build MILR upon a Transformer-based MUG model and choose the unified latent space represented by the intermediate output vectors. Since the shared latent space is modality-agnostic, it provides a unified view of multimodal reasoning, reducing the modality gap and improving the overall efficacy of cross-modal reasoning.

Reasoning with MILR involves finding the best vector representations of image and text tokens that lead to improved image quality. We implement it using the policy gradient method (Williams, 1992), where, at test time, the reward is computed by scoring the compatibility between the generated image and the given instruction. Crucially, gradients are only back-propagated to the cross-modal latent representations (i.e., the intermediate model outputs) obtained from the forward pass (see Figure 1), without altering any model parameters, and thus making MILR a test-time latent reasoning method.

We evaluate MILR on three widely-used image generation benchmarks: GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), and WISE (Niu et al., 2025). MILR established new state-of-the-art results across all benchmarks. Notably, it achieves an overall score of 0.95 on GenEval, matching the best training-based model and outperforming the best test-time-scaling method by 4.4%. On the more challenging WISE benchmark, MILR obtains an overall score of 0.63, surpassing the strongest baseline model by 16.7%. Our further analysis reveals that the best performance of MILR is driven by its ability to perform joint image-text reasoning in a unified latent space. This unique advantage also makes MILR successful in instructions that involve challenging cultural and temporal reasoning.

To summarize:

- We introduce MILR, a test-time reasoning-augmented method that improves image generation by performing joint image-text reasoning in a unified latent space.
- We demonstrate the effectiveness of MILR by showing that it achieves superior performance across three different benchmarks for image generation.
- We conduct a comprehensive analysis, perform various ablation studies, and discuss potential limitations of MILR.

2 RELATED WORKS

Reasoning-Augmented Image Generation. Remarkable progress has been achieved in reasoning-augmented LLMs (Wei et al., 2022; DeepSeek-AI et al., 2025), but effectively integrating reasoning into conventional text-guided image generation models remains a challenge (Betker et al., 2023; Sun et al., 2024; Esser et al., 2024; Zhao et al., 2024; Chen et al., 2025). With the development of unified multimodal understanding and generation that can generate images and text in an interleaved manner, many works have explored reasoning for text-guided image generation. The predominant paradigm uses causal language modeling to unify image and text generation as next token prediction, enabling the model to make plans via chain-of-thought before generating images (Fang et al., 2025; Deng et al., 2025; Xiao et al., 2025; Cai et al., 2025; Guo et al., 2025; Jiang et al., 2025a; Duan et al., 2025). Another line of research focuses on reasoning over images via the test-time scaling strategy (Li et al., 2025b; Zhuo et al., 2025). These works typically rely on an external critic model to provide feedback on further improvement. Unlike all these methods that explicitly perform reasoning over raw images and text, we use test-time optimization to refine the latent representations of image and text tokens, leading to a unified cross-modal latent reasoning method.

Reasoning in The Latent Space. Differently from explicit reasoning via a chain of thoughts (Wei et al., 2022), latent reasoning refers to an implicit reasoning mechanism applied to the latent states (e.g., intermediate outputs) of the model. As in Transformer-based models, latent reasoning is typically implemented as spatial and temporal recurrences. Spatial recurrences implicitly deepen the model by iterating over the latent states between the Transformer layers (Hao et al., 2024; Cheng & Van Durme, 2024; Zhang et al., 2025a; Shen et al., 2025), while temporal recurrences refine the latent states through iterations across input tokens (Dao & Gu, 2024; Geiping et al., 2025). Under the recurrences is the idea of scaling up test-time computation for inference; however, this requires the recurrent modules to be pre-trained. In a similar vein to those, we iterate over the unified image-text latent states only at test time, without introducing and updating any model parameters.

Reinforcement Learning for Reasoning. Reinforcement learning (RL) has been the key to eliciting the reasoning ability of large language models (DeepSeek-AI et al., 2025; OpenAI et al., 2024). Inspired by the success of GRPO, which is an improved RL algorithm over PPO (Schulman et al., 2017) and is used in DeepSeek-R1 (Shao et al., 2024), researchers have extended it to the multimodal understanding domain, including visual question answering (Huang et al., 2025b; Liu et al., 2025b). In image generation, RL has also been shown to be effective (Jiang et al., 2025a; Tong et al., 2025; Jiang et al., 2025b; Pan et al., 2025b; Duan et al., 2025; Pan et al., 2025a; Liu et al., 2025a; Xue et al., 2025; Xiao et al., 2025), but has been used as a training-time optimization method. Unlike them, we employ a simple algorithm, REINFORCE (Williams, 1992), for test-time optimization.

3 METHOD

3.1 REASONING-AUGMENTED MULTIMODAL IMAGE GENERATION

We are interested in a framework for reasoning-augmented image generation that natively supports language reasoning before image synthesis. A typical implementation is through unified multimodal understanding and generation (MUG; Chen et al. (2025); Deng et al. (2025)). Suppose the language token sequence $\mathbf{t} \coloneqq t_{1:M} \coloneqq t_1, t_2, \ldots, t_M$, the image token sequence $\mathbf{v} \coloneqq v_{1:N} \coloneqq v_1, v_2, \ldots, v_N$, and the given instruction c, MUG defines multi-modal image generation as an autoregressive generation process:

$$p(\mathbf{t}, \mathbf{v}|c) = \prod_{n=1}^{N} p(v_n|v_{1:n}, \mathbf{t}, c) \prod_{m=1}^{M} p(t_m|t_{1:m}, c)\,, \tag{1}$$

The generation of $\mathbf{v}$ depends on the reasoning via language tokens $\mathbf{t}$, which are generated from c. By analogy to textual reasoning, we refer to the generation of $\mathbf{v}$ as visual reasoning. To produce the final image, $\mathbf{v}$ is further passed through a pixel decoder: $V_f \sim p(\cdot|\mathbf{t}, \mathbf{v}, c)$. Since we focus on discrete image tokens produced by a pre-trained discrete VAE (Chen et al., 2025), the pixel decoder $p(\cdot|\mathbf{t}, \mathbf{v}, c)$ becomes deterministic given $\mathbf{v}$.

Let $R(V_f, c)$ denote a reward model that scores the compatibility between the given instruction c and the final image V_f, the goal of test-time reasoning-augmented image generation is to find an

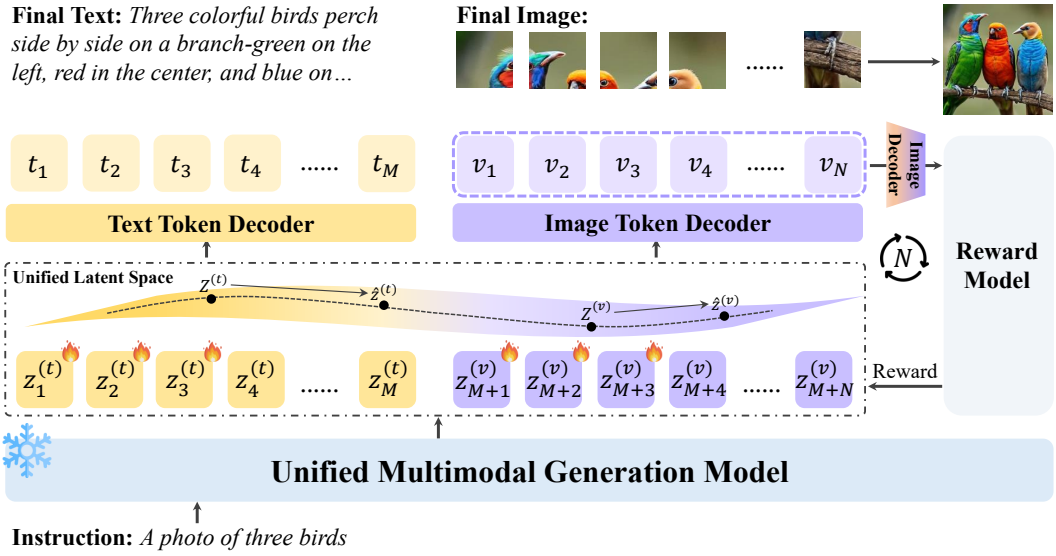


Figure 2: Overview of MILR. MILR performs test-time latent reasoning in a unified latent space; it uses policy gradients to iteratively refine text & image latents $\mathbf{z}^{(t)}, \mathbf{z}^{(v)}$, guided by a reward model. The reward model scores each generated image conditioning on the instruction.

optimal pair $(\mathbf{t}^*, \mathbf{v}^*)$ that maximizes the expected reward under $p(\cdot|\mathbf{t}, \mathbf{v}, c)$, without modifying any model parameters:

$$\mathbf{t}^*, \mathbf{v}^* = \arg\max_{\mathbf{t},\mathbf{v}} \mathbb{E}_{V_f \sim p(\cdot|\mathbf{t},\mathbf{v},c)}[R(V_f, c)] \,. \tag{2}$$

In the case of MUG, the model itself can act as the reward function because of its multimodal understanding ability; alternatively, any off-the-shelf model that has such a capability can serve the same role (see discussion in Section 4.4). Due to the infinite search space, the problem defined by Equation 2 is generally intractable.

3.2 Image Generation via Test-Time Latent Reasoning

3.2.1 Multi-Modal Latent Reasoning

Rather than searching over discrete image and text tokens, we propose searching in the unified latent vector space, that is, searching over their continuous vector representations. As in MUG, these vectors correspond to the intermediate model outputs at respective token positions. They lie in a vector space that encodes both image and text tokens, offering a unified view of visual and textual reasoning[1] and facilitating cross-modal reasoning (see Figure 2).

Formally, denoting the latent representations of image and text tokens by $\mathbf{z}^{(v)} = z^{(v)}_{1:N}$ and $\mathbf{z}^{(t)} = z^{(t)}_{1:M}$, respectively, where $z^{(v)}, z^{(t)} \in \mathbb{R}^d$ are the outputs from the same Transformer layer and thus lie in a shared d-dimensional vector space, we can rewrite Equation 2 as:

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} \mathbb{E}_{V_f \sim p(\cdot|\mathbf{z}.c)}[R(V_f, c)] \,, \tag{3}$$

where $\mathbf{z} = [\mathbf{z}^{(t)}; \mathbf{z}^{(v)}]$ indicates the multimodal latent representation of the token sequence $[\mathbf{t}, \mathbf{v}]$. We refer to this optimization problem as multimodal latent reasoning.

Given the optimal $\mathbf{z}^\star$ from a specific model layer, to produce the final V_f, we need to continue the forward pass until it is decoded into discrete tokens $[\mathbf{t}, \mathbf{v}]$. Thus, the pixel image generation becomes:

$$p(V_f|\mathbf{z}^*, c) = p(V_f|\mathbf{t}, \mathbf{v}, c)p(\mathbf{t}, \mathbf{v}|\mathbf{z}^*) \,, \tag{4}$$

where $p(\mathbf{t}, \mathbf{v}|\mathbf{z}^*)$ represents the remaining forward pass of MUG starting with $\mathbf{z}^*$.

[1] We refer to "reasoning" as iterative updates in the latent space rather than text-based chain-of-thought.

3.2.2 GRADIENT-BASED OPTIMIZATION FOR LATENT REASONING

In general, the problem defined by Equation 3 does not admit a closed-form solution, so we resort to REINFORCE (Williams, 1992), a policy gradient optimization method. We note that it has been applied to fully textual reasoning in language tasks (Li et al., 2025a), but we, for the first time, extend it to unified multimodal latent reasoning for image generation.

With REINFORCE, we can formulate the cross-modal optimization process as:

$$\mathbf{z}^{k+1} \leftarrow \mathbf{z}^{k} + \eta \cdot \mathcal{J}(\mathbf{z}^{k}), \tag{5}$$

$$\mathcal{J}(\mathbf{z}^{k}) = \mathbb{E}_{V_f \sim p(\cdot|\mathbf{z}^{k}, c)}[R(V_f, c)\nabla_{\mathbf{z}} \log(p(\mathbf{t}, \mathbf{v}|\mathbf{z}^{k}))]\,, \tag{6}$$

where η is the learning rate. For efficiency, we choose as $\mathbf{z}$ the outputs of the last Transformer layer, i.e., the inputs to the final modal-specific decoding heads. Moreover, we approximate gradients $\mathcal{J}(\mathbf{z})$ using a single sampled pair $(\mathbf{t}, \mathbf{v})$. The gradients are back-propagated only to the model outputs $\mathbf{z}$, without altering any model parameters, and thus making MILR a test-time reasoning method.

Naively, we would optimize[2] all $M + N$ latents in $z_{1:M+N}$, but searching using only the guidance of a reward model is potentially biased, and it does not leverage the generative capacity of MUG for better exploration; instead, we optimize only the first $\lambda_t M$ (where $\lambda_t \in (0, 1]$) latents for text. After decoding them into discrete tokens, we complete textual reasoning via the standard autoregressive generation conditioned on them. As we will later see, this simple strategy strikes a good balance between efficiency and performance (see Section 4.3.1). For visual reasoning, we adopt a similar strategy and optimize the first $\lambda_v N$ (where $\lambda_v \in (0, 1]$) latents.[3] This is further supported by the observations of Hu et al. (2025): the first few tokens govern the global structure of the image, while the remaining tokens primarily influence high-frequency details.[4]

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Benchmarks and Baselines. We conduct experiments on three benchmarks widely used in image generation: GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), and WISE (Niu et al., 2025). We compare MILR against the following three sets of baselines:

- **Non-reasoning models** refer to the classical image generation models that synthesize images from a given instruction in a single shot, without refining the instruction. We consider diffusion models such as FLUX.1-dev (Black Forest Labs, 2024), DALL·E 3 (Betker et al., 2023), and SD3-Medium (Esser et al., 2024), autoregressive models such as LlamaGen (Sun et al., 2024) and Emu3 (Wang et al., 2024), and hybrid autogressive-diffusion models such as BAGEL (Deng et al., 2025) and GPT-4o (OpenAI, 2025).
- **Training-based reasoning models** refer to the image generation models that acquire the reasoning ability through training, including GoT-R1 (Duan et al., 2025), T2I-R1 (Jiang et al., 2025a), Flow-GRPO (Liu et al., 2025a), and GRPO- and DPO-tuned Janus-Pro (Tong et al., 2025).
- **Test-time reasoning models** refer to the models that admit reasoning through tailored inference strategies, including Reflect-DiT (Li et al., 2025b) and ReflectionFlow (Zhuo et al., 2025), which uses language feedback, and Best-of-N and PARM (Guo et al., 2025), which rely on search.

Hyperparameters and configurations. We choose Janus-Pro (Chen et al., 2025), an autoregressive model, as the MUG model. For the portions of text and image tokens that are optimized, we perform grid search on a validation split sampled from GenEval and empirically set $\lambda_t = 0.2$ for text and $\lambda_v = 0.02$ for images (see Figure 5). We use the Adam optimizer (Kingma, 2014), where the learning rate is empirically set to 0.03. For each benchmark, we use its own evaluation toolkit as the reward model, following previous work (Liu et al., 2025a; Jiang et al., 2025a; Tong et al., 2025). We

[2] We experimented with different optimization strategies, see details in Appendix A.1.

[3] In each iteration, the text token number M may vary while the image token number N remains unchanged.

[4] Rather than optimizing a prefix of image token sequence, we can optimize a random subset of image tokens, but this strategy proves worse than prefix optimization (see Table 5 in Appendix A.1).

Table 1: Results on GenEval. The best score is in bold and the second best is underlined.

Method	Single Obj. ↑	Two Obj. ↑	Counting ↑	Colors ↑	Position ↑	Attr. Binding ↑	Overall ↑
Non-reasoning Models							
LlamaGen (Sun et al., 2024)	0.71	0.34	0.21	0.58	0.07	0.04	0.32
Emu3 (Wang et al., 2024)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
FLUX.1-dev (Black Forest Labs, 2024)	0.98	0.79	0.73	0.77	0.22	0.45	0.66
DALL-E 3 (Betker et al., 2023)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium (Esser et al., 2024)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
BAGEL (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82
GPT-4o (OpenAI, 2025)	0.99	0.92	0.85	0.91	0.75	0.66	0.85
Training-based Reasoning Models							
GoT-R1 (Duan et al., 2025)	0.99	0.94	0.50	0.90	0.46	0.68	0.75
T2I-R1 (Jiang et al., 2025a)	0.99	0.91	0.53	0.91	0.76	0.65	0.79
Flow-GRPO (Liu et al., 2025a)	**1.00**	**0.99**	**0.95**	0.92	**0.99**	0.86	**0.95**
ReasonGen-R1 (Zhang et al., 2025b)	0.99	0.94	0.62	0.90	0.84	0.84	0.86
Janus-Pro-7B(+GRPO) (Tong et al., 2025)	0.99	0.87	0.61	0.87	0.82	0.68	0.81
Janus-Pro-7B(+DPO) (Guo et al., 2025)	0.99	0.89	0.65	0.92	0.82	0.72	0.83
Test-time Reasoning Models							
Reflect-DiT (Li et al., 2025b)	0.98	0.96	0.80	0.88	0.66	0.60	0.81
ReflectionFlow (Zhuo et al., 2025)	**1.00**	0.98	0.90	0.96	0.93	0.72	0.91
Janus-Pro-7B(+Text Enhanced Reasoning)	0.98	0.91	0.55	0.89	0.74	0.67	0.79
Janus-Pro-7B(+Best-of-N)	0.99	0.96	0.89	0.93	0.92	0.80	0.91
Janus-Pro-7B(+PARM) (Guo et al., 2025)	**1.00**	0.95	0.80	0.93	0.91	0.85	0.91
Janus-Pro-1B (Chen et al., 2025)	0.98	0.82	0.51	0.89	0.65	0.56	0.73
Janus-Pro-1B+MILR	**1.00**	0.91	0.78	0.92	0.86	0.86	0.89
Janus-Pro 7B (Chen et al., 2025)	0.98	0.85	0.56	0.89	0.77	0.64	0.78
Janus-Pro-7B+MILR	**1.00**	0.96	0.90	**0.98**	0.98	**0.91**	**0.95**

Table 2: Results on T2I-CompBench and WISE. The best is in bold, and the second is in underlined.

Method	T2I-CompBench							WISE
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	Complex. ↑	Overall ↑	Avg ↑
Non-reasoning Models								
PixArt-α Chen et al. (2023)	0.6690	0.4927	0.6477	0.2064	0.3197	0.3433	0.4465	0.47
FLUX.1-dev Black Forest Labs (2024)	0.7407	0.5718	0.6922	0.2863	0.3127	0.3703	0.4957	0.50
DALL-E 3 Betker et al. (2023)	0.7785	**0.6209**	0.7036	0.2865	0.3003	0.3773	0.5112	-
SD3-Medium Esser et al. (2024)	0.8132	0.5885	0.7334	0.3200	0.3140	0.3771	0.5244	0.42
Show-o Xie et al. (2024)	0.5600	0.4100	0.4600	0.2000	0.3000	0.2900	0.3700	0.30
BAGEL Deng et al. (2025)	0.8027	0.5685	0.7021	0.3488	0.3101	0.3824	0.5191	0.52
Training-based Reasoning Models								
T2I-R1 Jiang et al. (2025a)	0.8130	0.5852	0.7243	0.3378	0.3090	0.3993	0.5281	0.54
GoT-R1 Duan et al. (2025)	0.8139	0.5549	**0.7339**	0.3306	0.3169	**0.3944**	0.5241	-
Janus-Pro-7B(+GRPO) Tong et al. (2025)	0.7721	0.5366	0.7317	0.2869	0.3087	0.3697	0.5010	-
Test-time Reasoning Models								
Show-o + PARM Guo et al. (2025)	0.7500	0.5600	0.6600	0.2900	0.3100	0.3700	0.4900	-
Janus-Pro-7B(+Text Enhanced Reasoning)	0.7087	0.4419	0.5821	0.2597	0.3072	0.3761	0.4459	0.46
Janus-Pro-7B(+Best-of-N)	0.7089	0.4925	0.7089	0.3542	**0.3262**	0.3721	0.4938	0.52
Janus-Pro-1B Chen et al. (2025)	0.3411	0.2261	0.2696	0.0968	0.2808	0.2721	0.2478	0.26
Janus-Pro-1B+MILR	0.6066	0.2796	0.4177	0.2796	0.2622	0.2613	0.3512	0.40
Janus-Pro-7B Chen et al. (2025)	0.6359	0.3528	0.4936	0.2061	0.3085	0.3559	0.3921	0.35
Janus-Pro-7B+MILR	**0.8508**	0.5117	0.6949	**0.4613**	0.3078	0.3684	**0.5325**	**0.63**

conduct all experiments on a single NVIDIA A100 80GB GPU. A discussion on model efficiency can be found in Appendix A.3.3.

4.2 Main Results

MILR achieves state-of-the-art results on GenEval, one of the most widely-used benchmarks for image generation (see Table 1). It improves over the base Janus-Pro-7B by 0.17, with the largest increases obtained from *Counting* (+0.34), *Position* (+0.21), and *Attribute Binding* (+0.27). Notably, MILR surpasses frontier non-reasoning models such as SD3-Medium, BAGEL, and GPT-4o (+12%). Compared with training-based reasoning models (e.g., GoT-R1 and T2I-R1), MILR performs better and requires no parameter tuning. For fairness, we also compare MILR with test-time reasoning models. Surprisingly, it outperforms ReflectionFlow and PARM (+4.5%) that rely on scaling up test-time computation, demonstrating the superiority of our test-time optimization method.

We further evaluate MILR on two additional benchmarks: T2I-CompBench and WISE (see Table 2). Again, it achieves the best performance on both, highlighting the robustness of our model. Specifically, on T2I-CompBench, MILR improves over the base Janus-Pro-7B by a large margin (+0.14)

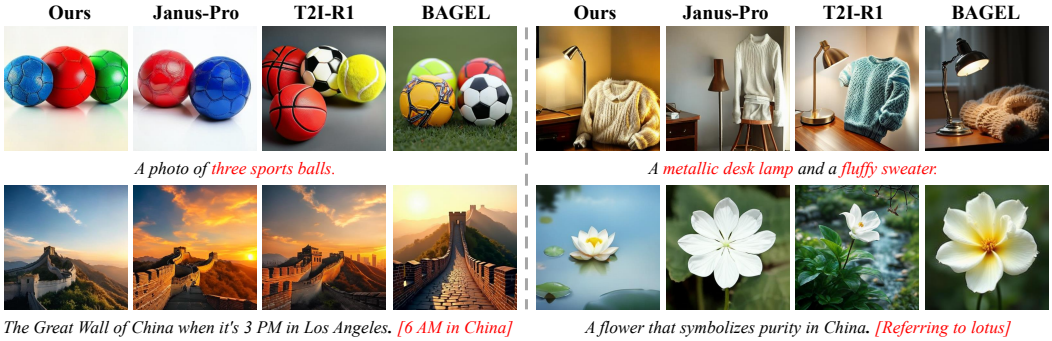


Figure 3: Qualitative studies on theree benchmarks. Reasoning cues are highlighted in red.

Table 3: Ablations of MILR on GenEval, T2I-CompBench, and WISE

Method	GenEval							T2I-CompBench	WISE
	Single Obj.	Two Obj.	Counting	Color	Pos.	Attr. Binding	Overall	Overall	Avg
Janus-Pro-7B+MILR (ours)	1.00	0.96	0.90	0.98	0.98	0.91	**0.95**	**0.5325**	**0.63**
w/o MILR	0.98	0.85	0.56	0.89	0.77	0.64	0.78	0.3921	0.35
w/o Image	1.00	1.00	0.91	0.95	0.95	0.88	0.94	0.5210	0.61
w/o Text	1.00	0.95	0.88	0.91	0.97	0.89	0.93	0.5043	0.56

and slightly outperforms T2I-R1, a strong training-based reasoning model. On WISE, which emphasizes world knowledge understanding, MILR outperforms the base Janus-Pro-7B (+80%) and the second-best model T2I-R1 (+16.7%), implying the importance of reasoning in comprehending knowledge-intensive instructions (see our qualitative studies of reasoning trajectories in Figure 8).

MILR is capable of geometric, temporal, and cultural reasoning. Next, we conduct a qualitative study on knowledge-intensive WISE. Surprisingly, MILR demonstrates nontrivial ability in geometric, temporal and cultural reasoning (see Figure 3). Take the prompt "The Great Wall of China when it's 3 PM in Los Angeles", MILR correctly infers the time difference from its geometric knowledge of China and Los Angeles, and concludes that "The Great Wall at Dawn" is of interest. As another example, MILR correctly infers that lotus symbolizes purity in Chinese culture.

Joint image-text reasoning in the unified latent space leads to the best performance. To better understand the contribution of each modality to the strong performance of MILR, we individually ablate the latent optimization of images and text, denoted by "w/o image" and "w/o text", respectively (see Table 3). First, we find that both of them exceed the base model (w/o MILR) by a large margin (e.g., > 0.21 on WISE), and optimizing both modalities leads to the best performance. Interestingly, text-only optimization (w/o image) fares slightly better than image-only optimization (w/o text), and approaches the performance of our best model (+MILR) on all benchmarks, suggesting substantial room for improvement in the language understanding component of MUG-based image generation models.

4.3 Analysis

4.3.1 Hyperparameters

We analyze three important hyperparameters of MILR: (1) the maximum optimization step T, (2) the portion of text tokens to be optimized λ_t in text-only optimization, and (3) the proportion of image tokens to be optimized λ_v in image-only optimization. We perform analysis 1 on all three benchmarks and analyses 2 and 3 on a held GenEval validation split. For each setting, we run MILR three times with different random seeds and report the average scores.

Scaling up the number of optimization steps improves performance on all benchmarks. MILR admits test-time compute scaling by increasing the number of optimization steps. We illustrate and compare three setups: MILR, text-only optimization, and image-only optimization (see Figure 4). On all benchmarks, increasing the number of optimization steps leads to consistent improvements, with the best performance achieved at step 16, after which model performance plateaus. Moreover, among the three setups, optimizing both images and text (i.e., MILR) yields the best per-

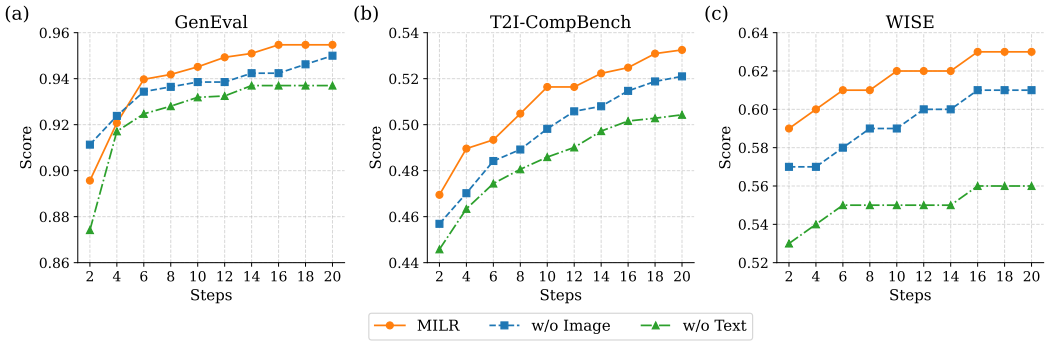


Figure 4: Performance across three benchmarks for varying optimization steps.

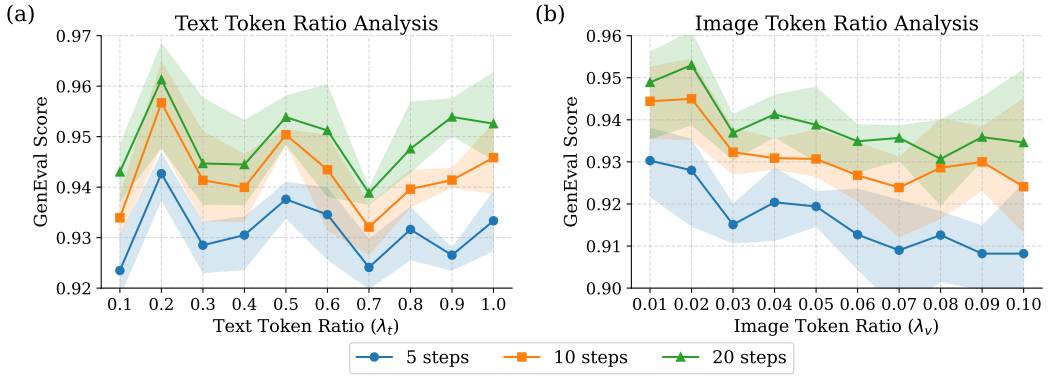


Figure 5: GenEval scores with varying optimization ratios of text and image tokens.

formance for almost all steps, except for a few early steps on GenEval, suggesting that MILR is well-suited for test-time compute scaling.

Optimizing a moderate amount (e.g., 20%) of text tokens leads to the best performance. When varying λ_t from 0.1 to 1.0, text-only optimization fluctuates between 0.91 and 0.96, and reaches the highest score at $\lambda_t = 0.2$ (see Figure 5 (a)). We observe the same trend for different numbers of optimization steps. All of these suggest that MILR is relatively robust to text optimization. Moreover, we empirically find that the optimal $\lambda_t = 0.2$ corresponds to the prompts that elicit more coherent reasoning (see examples in Figure 10).

Optimizing a tiny amount (e.g., 2%) of image tokens gives rise to the best result. Hu et al. (2025) have empirically shown that, in autoregressive image generation, early-stage tokens govern the overall image structure, and perturbing the first 20% of image tokens results in significant structural deviations. Therefore, we conservatively adopt a very small λ_v, varying it from 0.01 to 0.1 (see Figure 5 (b)). Surprisingly, optimizing only the first 2% of all image tokens already yields peak performance, while further increasing λ_v tends to degrade it (see examples in Figure 12).

4.4 Reward Models

The reward model is a crucial component of MILR, providing learning signals for latent reasoning. Following previous work (Liu et al., 2025a; Tong et al., 2025), we have used the benchmark's evaluator as the reward model (denoted by 👑 *OracleReward*), but in real-world scenarios, oracle rewards are usually unknown, and it is difficult to design domain-specific rewards. To show that MILR is effective without the reliance on 👑, we test it with a set of off-the-shelf reward models on GenEval:

- 💫 *SelfReward* uses MUG itself (e.g., Janus-Pro in this work) to evaluate images.
- *GPT-4o* represents a frontier critic for assessing image quality (Hurst et al., 2024).
- 🔗 *UnifiedReward* is specifically tuned for a unified evaluation of MUG (Wang et al., 2025).

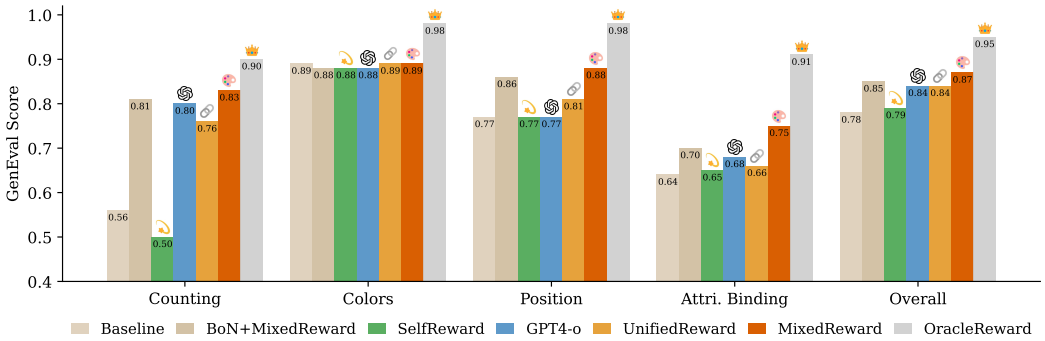


Figure 6: Performance with different reward models on GenEval.

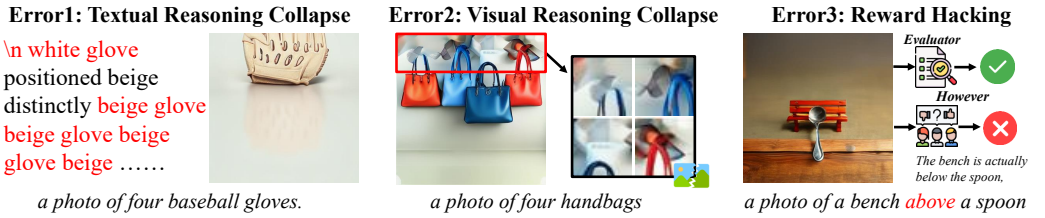


Figure 7: Error Case Study

- *MixedReward* is a composite critic for more comprehensive evaluation. It aggregates rewards from specialized models, including GroundingDINO (Liu et al., 2024) (evaluating object detection), GIT (Wang et al., 2022) (judging colors), and (assessing aesthetics).

Unsurprisingly, gives rise to the best performance across all dimensions (see Figure 6). For non-oracle critics, all variants surpass the baseline in terms of the overall score. Notably, MILR remains relatively robust to different reward models, except for , which performs poorly on *Counting* (around 0.5). Among non-oracle critics, performs the best, suggesting that, in the absence of oracle rewards, we can derive a strong universal reward model by combining specialized critic models. Moreover, MILR+ slightly outperforms the strong Best-of-N+ baseline (+2.4%) under comparable computation (i.e., N = T = 20), once again demonstrating the superiority of our method.

4.4.1 Error Cases

Though MILR has shown the best performance across all three benchmarks, we observe three major failure modes and illustrate them respectively in Figure 7. Specifically, the failure modes include (1) **Textual Reasoning Collapse.** The regenerated chain of thoughts degenerates into repetitive phrases and becomes nonsensical for guiding image generation (e.g., repeated "beige glove"). (2) **Visual Reasoning Collapse.** While textual reasoning is fine, visual reasoning degrades prefix image tokens that govern the overall structure and subsequently adversely affects the generation of remaining tokens that control fine-grained details (e.g., blurred handles). (3) **Reward Hacking.** Models exploit shortcuts to achieve high rewards, but the generated image does not align with the given instruction perfectly (e.g., unmatched position relationship). This implies that the benchmark's evaluator is limited in that it is not yet fully capable of spatial reasoning.

5 Discussion

We acknowledge two primary limitations of MILR. First, we have focused on an implementation of MILR that builds upon autoregressive MUG, where both text and image tokens are generated autoregressively. Another strong paradigm of MUG is through diffusion image generation, where image tokens within the same image can attend to each other (Deng et al., 2025). It is interesting to see if the effectiveness of MILR transfers. Second, MILR relies on a reward model for learning signals, but, in practice, a perfect reward model usually does not exist, and it is difficult to design a domain-agnostic reward model. In addition, our experimental results have revealed that the strongest non-oracle reward model still lags behind the oracle reward. Thus, future work is well-suited for designing reward models that can generalize like the unified reward model of Wang et al. (2025).

6 CONCLUSION

We have proposed MILR, a test-time latent reasoning method for multimodal image generation. MILR employs policy gradient optimization, guided by an image quality metric, to search over the latent vector representations of discrete image and text tokens, leading to a unified multimodal reasoning framework. We implement MILR with a unified multimodal understanding and generation model that natively supports language reasoning before image synthesis. Across three benchmarks, MILR achieves state-of-the-art results. We further perform an in-depth analysis; we find that jointly reasoning in the latent image-text space is the key to its strong performance. Our qualitative studies also highlight MILR's nontrivial capability in temporal and cultural reasoning.

7 ETHICS STATEMENT

We strictly adhere to the ICLR Code of Ethics. Below we elaborate on the ethical considerations relevant to our work and the measures we have put in place.

Safety of content. To reduce potential risks, we confine our text prompts to those of the standard benchmarks, avoiding sensitive and offensive input. We manually reviewed the generated images and text used in the work and did not find harmful content.

Privacy and intellectual property. Our work does not involve processing personally identifiable information, as we experiment exclusively on publicly available benchmarks and off-the-shelf models, as per their respective licenses.

8 REPRODUCIBILITY STATEMENT

To ensure reproducibility, we present experimental details, including settings and hyperparameter configurations, in Section 4.1 and release our code at https://github.com/spatigen/milr.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf, 2(3):8, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Miaomiao Cai, Guanjie Wang, Wei Li, Zhijun Tu, Hanting Chen, Shaohui Lin, and Jie Hu. Autoregressive image generation with vision full-view prompt. *arXiv preprint arXiv:2502.16965*, 2025.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *International Conference on Learning Representations (ICLR)*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images, 2025. URL https://arxiv.org/abs/2505.22525.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *International Conference on Machine Learning (ICML)*, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

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

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025.

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.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

Jonas Geiping, Sean Michael McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*, 2025. URL https://openreview.net/forum?id=D6o6Bwtq7h.

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.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Taihang Hu, Linxuan Li, Kai Wang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. Anchor token matching: Implicit structure locking for training-free ar image editing, 2025. URL https://arxiv.org/abs/2504.10434.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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*, 2025a.

Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.17534*, 2025b.

Diederik P Kingma. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2014.

Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, and Zilong Zheng. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space, 2025a. URL https://arxiv.org/abs/2505.13308.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. *International Conference on Computer Vision (ICCV)*, 2025b.

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*, 2025a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *International Conference on Computer Vision (ICCV)*, 2025b.

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.

OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025. Accessed: 2025-07-25.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Jiadong Pan, Zhiyuan Ma, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Self-reflective reinforcement learning for diffusion-based image reasoning generation. *arXiv preprint arXiv:2505.22407*, 2025a.

Kaihang Pan, Wendong Bu, Yuruo Wu, Yang Wu, Kai Shen, Yunfei Li, Hang Zhao, Juncheng Li, Siliang Tang, and Yueting Zhuang. Focusdiff: Advancing fine-grained text-image alignment for autoregressive visual generation through rl. *arXiv preprint arXiv:2506.05501*, 2025b.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025.

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

Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

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.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Huimin Wu, Xiaojian Ma, Haozhe Zhao, Yanpeng Zhao, and Qing Li. Nep: Autoregressive image editing via next editing token prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Mingrui Wu, Lu Wang, Pu Zhao, Fangkai Yang, Jianjin Zhang, Jianfeng Liu, Yuefeng Zhan, Weihao Han, Hao Sun, Jiayi Ji, Xiaoshuai Sun, Qingwei Lin, Weiwei Deng, Dongmei Zhang, Feng Sun, Qi Zhang, and Rongrong Ji. Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning, 2025b. URL https://arxiv.org/abs/2505.17540.

Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo. *arXiv preprint arXiv:2505.13031*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *International Conference on Learning Representations (ICLR)*, 2024.

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.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*, 2025a.

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, 2025b. URL https://arxiv.org/abs/2505.24875.

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. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. *International Conference on Computer Vision (ICCV)*, 2025.

A Appendix

A.1 Optimization algorithms

By default, MILR jointly optimizes text latent $\mathbf{z}^{(t)}$ and image latent $\mathbf{z}^{(v)}$ in each gradient descent step, which is denoted as MILR-Joint. As an alternative, we can optimize $\mathbf{z}^{(t)}$ and $\mathbf{z}^{(v)}$ in an alternating fashion (denoted by MILR-Alt), akin to the strategy used in coordinate descent. We further propose another variant of MILR that first optimizes $\mathbf{z}^{(t)}$ to its approximate optimum, followed by optimizing $\mathbf{z}^{(v)}$ to its approximate optimum (denoted by MILR-T2V). We illustrate the three algorithms in Algorithm 1. Their performance on GenEval is presented in Table 4. We do not see significant differences among them, so we use MILR-Joint by default because of its simplicity.

Algorithm 1 MILR

Require: Instruction c, Learning rate η, MUG model p, reward threshold τ, text and image fraction $\lambda_t, \lambda_v \in (0, 1]$, optimization steps K, Reasoning strategy $\mathcal{S} \in \{\text{JOINT}, \text{ALT}, \text{T2V}\}$

$\mathbf{z}, \mathbf{t}, \mathbf{v} \leftarrow p(\mathbf{t}, \mathbf{v}|c)$ ▷ Initial latent vectors
$V_f \sim p(\cdot|\mathbf{v})$
$r \leftarrow R(V_f, c)$ ▷ Reward Calculation
$\mathbf{z}^0 \leftarrow (z^{(t)}_{1:\lambda_t|\mathbf{t}|}; z^{(v)}_{1:\lambda_v|\mathbf{v}|})$ ▷ Set λ_t and λ_v fraction
$k \leftarrow 1$ ▷ Starting step index
while $k \leq K$ and $r \leq \tau$ **do**
  $V_f, \mathbf{z}^k \leftarrow \mathit{LatentReasoning}_{\mathcal{S}}\left(\mathbf{z}^{k-1}, \eta, k, K, c, p\right)$ ▷ Select a strategy: Algorithm 2
  $r \leftarrow R(V_f, c)$
  $k \leftarrow k + 1$
end while
return V_f

Algorithm 2 LatentReasoning (default: JOINT)

(a) JOINT

$\mathbf{z}^k \leftarrow \mathbf{z}^{k-1} + \eta \cdot \mathcal{J}(\mathbf{z}^{k-1})$
$V_f \sim p(V_f \mid \mathbf{z}^k, c)$
return $V_f, \mathbf{z}^k$

(b) ALT

if $k \bmod 2 = 1$ **then**
  $\mathbf{z}^{(t)}_k \leftarrow \mathbf{z}^{(t)}_{k-1} + \eta \cdot \mathcal{J}(\mathbf{z}^{(t)}_{k-1})$
  $T_f \sim p(T_f|\mathbf{z}^{(t)}_k, c)$
  $V_f \sim p(V_f|\mathbf{z}^{(v)}_{k-1}, T_f, c)$
else
  $\mathbf{z}^{(v)}_k \leftarrow \mathbf{z}^{(v)}_{k-1} + \eta \cdot \mathcal{J}(\mathbf{z}^{(v)}_{k-1})$
  $V_f \sim p(V_f|\mathbf{z}^{(v)}_k, T_f, c)$
end if
return $V_f, \mathbf{z}^k$

(c) T2V

if $k \leq \lfloor K/2 \rfloor$ **then**
  $\mathbf{z}^{(t)}_k \leftarrow \mathbf{z}^{(t)}_{k-1} + \eta \cdot \mathcal{J}(\mathbf{z}^{(t)}_{k-1})$
  $T_f \sim p(T_f|\mathbf{z}^{(t)}_k, c)$
  $V_f \sim p(V_f|\mathbf{z}^{(v)}_{k-1}, T_f, c)$
else
  $\mathbf{z}^{(v)}_k \leftarrow \mathbf{z}^{(v)}_{k-1} + \eta \cdot \mathcal{J}(\mathbf{z}^{(v)}_{k-1})$
  $V_f \sim p(V_f|\mathbf{z}^{(v)}_k, T_f, c)$
end if
return $V_f, \mathbf{z}^k$

Table 4: Geneval results with different reasoning strategies.

Algorithms	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding	**Overall**
MILR-JOINT	1.00	0.96	0.90	0.98	0.98	0.91	0.95
MILR-ALT	1.00	0.98	0.87	0.95	0.97	0.88	0.94
MILR-T2V	1.00	0.96	0.95	0.95	0.96	0.89	0.95

Table 5: Geneval results under random subset optimization of image tokens.

Algorithms	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding	**Overall**
MILR (Image+Prefix)	1.00	0.95	0.88	0.91	0.97	0.89	0.93
MILR (Image+Random)	1.00	0.92	0.84	0.89	0.94	0.86	0.90

A.2 Qualitive Study

A.2.1 Reasoning trajectories

MILR admits multi-round latent-space refining at test time. In Figure 8, we visualize the full trajectories of rounds 2, 3, and 4 that finally arrive at the correct generation.

Figure 8: Example reasoning trajectories. Orange indicates prefix optimization, red highlights incorrect reasoning, and green represents correct reasoning.

A.2.2 Model Ablations

We provide a qualitative study accompanying the model ablations done in Section 4.2. Consistent with the quantitative results, single-modality optimization surpasses the base model but underperforms joint optimization. Clearly, text-only optimization excels at tasks that rely on nontrivial numerical and compositional reasoning (e.g., "a photo of four clocks/bowls/knives" or "three kites"), while image-only optimization tends to refine image structures by adjusting the spatial arrangement of objects (e.g., "a photo of a tie above a sink"). To illustrate this scenario, consider the prompt "a photo of four clocks" in Figure 9, text-only optimization produces four spatially uncorrelated clocks,

whereas image-only optimization tends to piece together four visually similar clocks on the same wall. Interestingly, across all examples, image-only optimization exhibits a tendency to zoom in on the objects of interest.

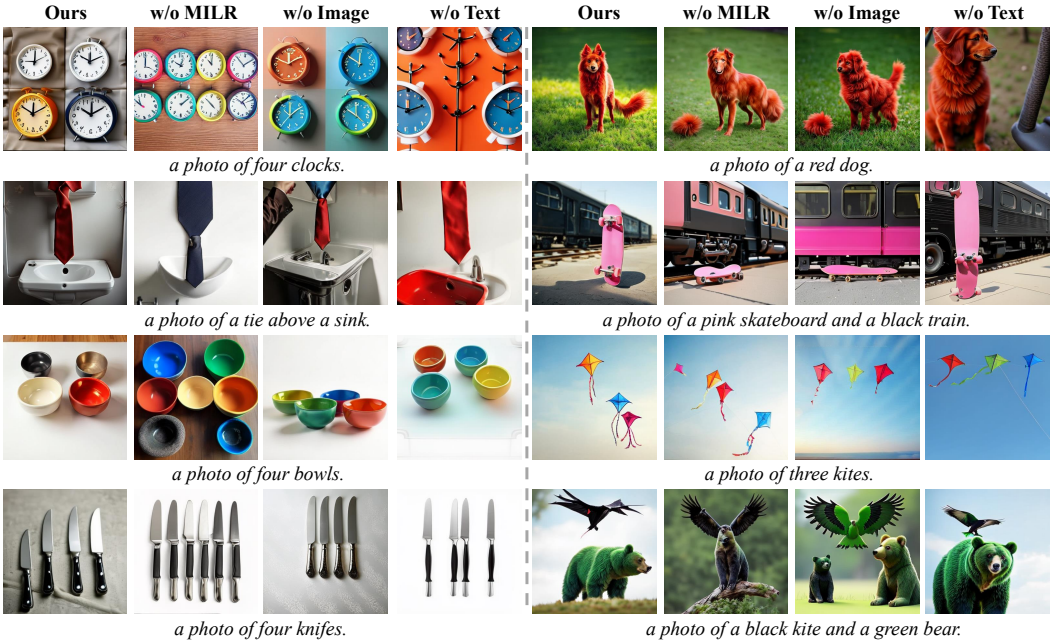


Figure 9: Example generations of MILR (ours), the base model (w/o MILR), text-only (w/o Image) optimization, and image-only (w/o Text) optimization.

A.3 Additional Analysis

A.3.1 Analysis of Text and Image Optimization Ratios

In Section 4.3.1, we studied how the text-token ratio λ_t and the image-token ratio λ_v affect MILR. We also provide example language and image reasoning trajectories when varying λ_t in Figure 10 and Figure 11, respectively. In Figure 12, we show how image generation evolves when varying λ_v. We empirically find that $\lambda_t = 0.2$ and $\lambda_v = 0.02$ generally lead to better generation.

A.3.2 Analysis of Reward Models

In Section 4.4, we evaluated MILR with different reward models: 💫 SelfReward, GPT-4o, 🔗 UnifiedReward, 🎨 MixedReward, and 👑 OracleReward (see Table 6, Table 7 and Table 8). We further provide example images generated under each setup in Figure 13. Overall, 👑 establishes an upper bar of generation quality, and 🎨, among non-oracle reward models, performs best because it incorporates multiple specialized critics to provide a comprehensive assessment. Though 🎨 still lags behind 👑, in the absence of 👑, we can compose a strong universal reward model from specialized off-the-shelf image critics.

Table 6: GenEval results with different rewards. The best performance is in bold.

Reward	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding	**Overall**
Baseline	0.98	0.85	0.56	0.89	0.77	0.64	0.78
💫 SelfReward	**1.00**	0.90	0.50	0.88	0.77	0.65	0.79
GPT-4o	0.98	**0.94**	0.80	0.88	0.77	0.68	0.84
🔗 UnifiedReward	**1.00**	0.92	0.76	0.89	0.81	0.66	0.84
🎨 MixedReward	**1.00**	0.90	**0.83**	**0.89**	**0.88**	**0.75**	**0.87**
👑 OracleReward	1.00	0.96	0.90	0.98	0.98	0.91	0.95

Prompt: a photo of an apple and a donut.

Initial Text Reasoning: The apple is red and round, with a smooth texture. The donut is brown, round, and has a slightly rough texture. They are placed next to each other on a white plate.

Latent Optimization:

$\lambda_t = 0.1$	*Visual vibrant depicts a* orange apple and a chocolate donut. *<eos>*
$\lambda_t = 0.2$	*Visual vibrant depicts a orange with don donut displays* a red apple and a golden brown donut. *<eos>*
$\lambda_t = 0.3$	*Visual vibrant depicts a orange with don donut displays\n image has round with* a hole in the center, and a bright red apple with a smooth skin. *<eos>*
$\lambda_t = 0.4$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green* apple. *<eos>*
$\lambda_t = 0.5$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green, and It appleuts* *<eos>*
$\lambda_t = 0.6$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green, and It appleuts has golden and with with* donut. *<eos>*
$\lambda_t = 0.7$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green, and It appleuts has golden and with with with covered powdered crumb raised* . *<eos>*
$\lambda_t = 0.8$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green, and It appleuts has golden and with with with covered powdered crumb raised surface with The sit next* to each other. *<eos>*
$\lambda_t = 0.9$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green, and It appleuts has golden and with with with covered powdered crumb raised surface with The sit next side to one other* . *<eos>*
$\lambda_t = 1.0$	*Visual vibrant depicts a orange with don donut displays\n image has round with has with while smooth green, and It appleuts has golden and with with with covered powdered crumb raised surface with The sit next side to one other. what plate surface with The* apple and donut are placed on. *<eos>*

Figure 10: Case study of the text token optimization ratio λ_t. The underlined text is decoded from the optimized latents. *<eos>* denotes the end-of-sentence token.

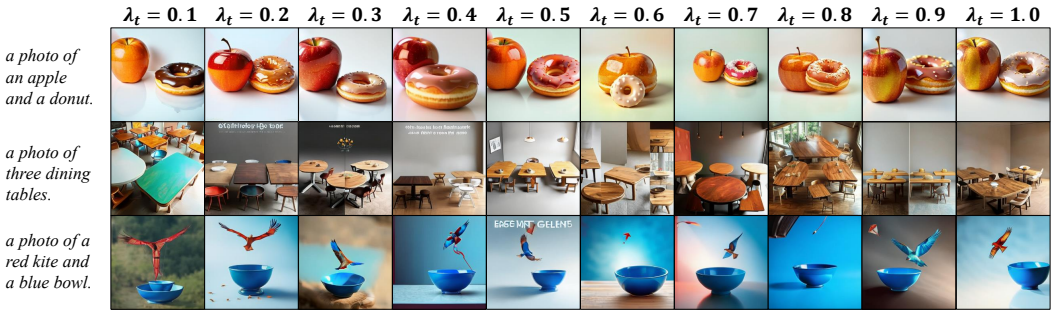


Figure 11: Generated images with varying token optimization ratio λ_t.

Table 7: T2I-CompBench results with different reward models. The best performance is in bold.

Reward	Color	Shape	Texture	Spatial	Non-Spatial	Complex	**Overall**
Baseline	0.6359	0.3528	0.4936	0.2061	0.3085	0.3559	0.3921
Self Reward	0.7196	0.4620	0.5887	0.2379	0.3055	0.3868	0.4501
GPT-4o	0.7624	0.4701	0.6561	0.2778	**0.3102**	**0.3881**	0.4775
UnifiedReward	0.8208	0.4609	0.5835	0.3210	0.3044	0.3700	0.4651
MixedReward	0.8009	0.4765	0.6608	0.4077	0.3055	0.3745	0.5043
OracleReward	**0.8508**	**0.5117**	**0.6949**	**0.4613**	0.3078	0.3684	**0.5325**

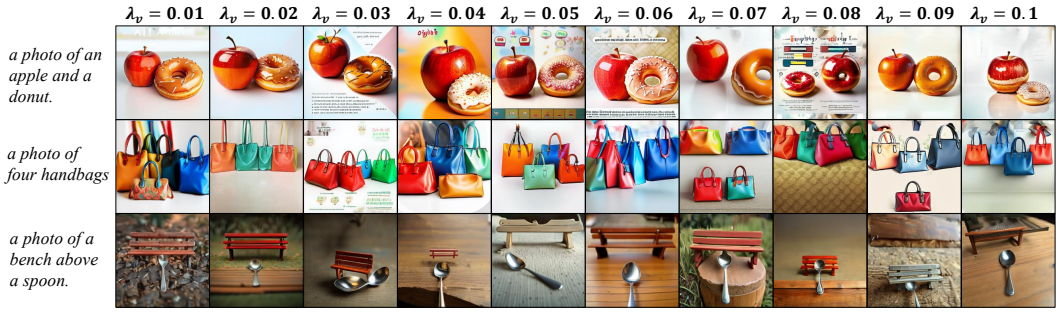


Figure 12: Generated images with varying image token optimization ratio λ_v.

Table 8: WISE results with different reward models. The best performance is in bold.

Reward	Cultural	Time	Space	Biology	Physics	Chemistry	**Overall**
Base	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Self Reward	0.40	0.43	0.53	0.38	0.44	0.23	0.41
GPT-4o	0.53	0.53	0.62	0.52	0.53	0.34	0.52
UnifiedReward	0.45	0.54	0.60	0.46	0.52	0.30	0.48
MixedReward	0.48	0.51	0.61	0.44	0.51	0.33	0.49
Metric Reward	**0.64**	**0.65**	**0.72**	**0.66**	**0.71**	**0.37**	**0.63**

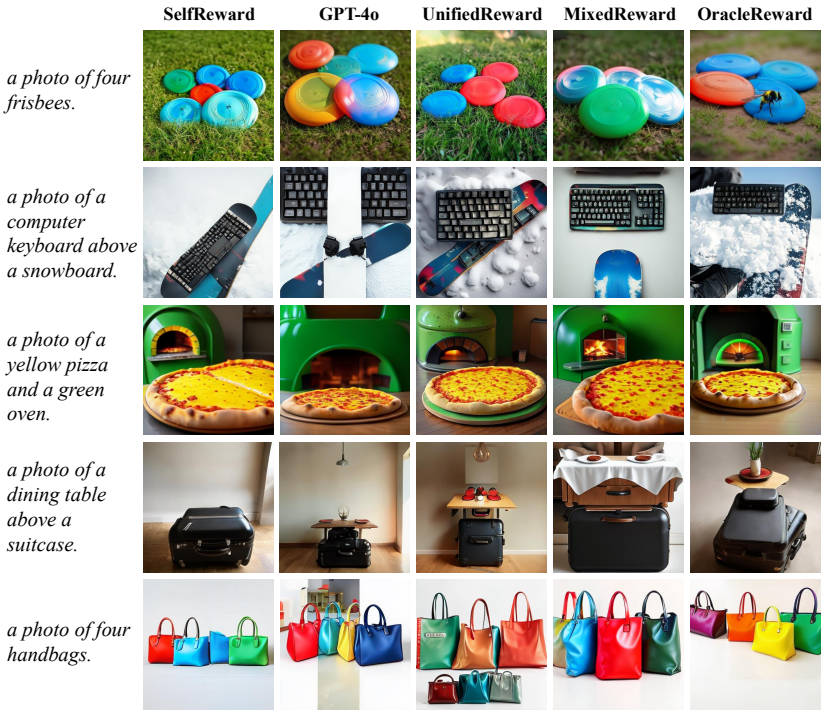


Figure 13: Case study of various reward models.

A.3.3 Efficieny Analysis

We summarize the computing resources used by different image generation models in Table 9. Compared with training-based reasoning approaches such as T2I-R1 (Jiang et al., 2025a), Janus-Pro with

DPO/GRPO (Tong et al., 2025), and Flow-GRPO (Liu et al., 2025a), our test-time reasoning method achieves the best performance, without relying on curated reasoning data for training. As an example, Flow-GRPO requires about 2K A800 GPU hours for training but only matches MILR, not to mention its labor cost in curating reasoning data. Moreover, when compared with the strong test-time reasoning method Best-of-N (N=20 vs. T=20 in MILR), MILR surpasses it by 0.04 while requiring less inference time because it employs an early-stop strategy, i.e., it stops once the generated image satisfies evaluation criteria.

Table 9: Efficiency comparisons.

Method	#GPU ↓	GPU	Training Time ↓	Inference Time ↓	Training Data	GenEval Score ↑
T2I-R1 (Jiang et al., 2025a)	8	H800	16 h	–	✓	0.79
Janus-Pro+GRPO (Tong et al., 2025)	8	A100	∼9 h	–	✓	0.81
Janus-Pro+DPO (Tong et al., 2025)	8	A100	∼9 h	–	✓	0.83
Flow-GRPO (Liu et al., 2025a)	24	A800	∼100 h	–	✓	**0.95**
Reflect-DiT (Li et al., 2025b)	–	A6000	24 h	16 h	✓	0.81
Janus-Pro-7B(+Best-of-N, N=20)	1	A100	0	8 h	✗	0.91
MILR (ours)	1	A100	0	5 h	✗	**0.95**

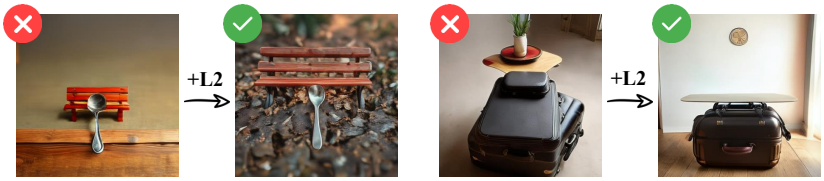
A.4 Latent Regularization to Mitigate Reward Hacking

As discussed in Section 4.4.1, MILR can suffer from reward hacking, which refers to a scenario where a MUG model achieves high rewards but exhibits low image generation quality. To mitigate this issue, we adopt a simple ℓ_2 regularizer to constrain the difference between the latents before and after optimization. Let $\mathbf{z}_{\text{init}}$ denote the initial latents produced by a frozen MUG model before optimization, and let $\mathcal{J}(\mathbf{z}^k)$ denote the original objective used at the latents optimization step k (see Equation (6)), we define the regularized objective as:

$$\mathcal{J}_{\text{reg}}(\mathbf{z}^k) = \mathcal{J}(\mathbf{z}^k) - \beta \left\| \mathbf{z}^k - \mathbf{z}_{\text{init}} \right\|_2^2, \tag{7}$$

where $\beta > 0$ controls the strength of regularization. Intuitively, this term encourages the optimized latents $\mathbf{z}^k$ to stay close to the initial latents, thus maintaining image generation capability similar to that of the frozen MUG model.

In our experiments, we use a small coefficient $\beta = 0.05$, which empirically stabilizes the optimization and alleviates semantic collapse, without degrading generation quality. In Figure 14, we illustrate that adding latent regularization helps mitigate reward hacking.



a photo of a bench above a spoon *a photo of a dining table above a suitcase.*

Figure 14: ℓ_2 regularization on latents mitigates reward hacking.

A.5 The Use of Large Language Models (LLMs)

We used LLMs (e.g., OpenAI ChatGPT) only to polish our writing and to create LaTeX algorithms and math equations. LLMs did not contribute to our research idea, model design, experimentation, etc. All contents of the paper have been verified by the authors.