T2I-R1: Reinforcing Image Generation with Collaborative Semantic-level and Token-level CoT

Dongzhi Jiang[*1], **Ziyu Guo**[*2], **Renrui Zhang**[*†1✉], **Zhuofan Zong**[1], **Hao Li**[3]
Le Zhuo[1,3], **Shilin Yan**, **Pheng-Ann Heng**[2], **Hongsheng Li**[1,3,4✉]

[1]CUHK MMLab [2]CUHK IMIXR [3]Shanghai AI Laboratory [4]CPII under InnoHK

{dzjiang, ziyuguo, renruizhang}@link.cuhk.edu.hk
hsli@ee.cuhk.edu.hk

[*]Equal Contribution [†]Project Leader [✉]Corresponding author

Abstract

Recent advancements in large language models have demonstrated how chain-of-thought (CoT) and reinforcement learning (RL) can improve performance. However, applying such reasoning strategies to the visual generation domain remains largely unexplored. In this paper, we present **T2I-R1**, a novel reasoning-enhanced text-to-image generation model, powered by RL with a bi-level CoT reasoning process. Specifically, we identify two levels of CoT that can be utilized to enhance different stages of generation: (1) the semantic-level CoT for high-level planning of the prompt and (2) the token-level CoT for low-level pixel processing during patch-by-patch generation. To better coordinate these two levels of CoT, we introduce **BiCoT-GRPO** with an ensemble of generation rewards, which seamlessly optimizes both generated CoTs within the same training step. By applying our reasoning strategies to the baseline model, Janus-Pro, we achieve superior performance with 13% improvement on T2I-CompBench and 19% improvement on the WISE benchmark, even surpassing the state-of-the-art model FLUX.1. All the training code and data are available at https://github.com/CaraJ7/T2I-R1.

1 Introduction

The emergence of advanced Large Language Models (LLMs) [62, 64, 81, 95], such as OpenAI o1 [65] and DeepSeek-R1 [20], has demonstrated considerable reasoning capabilities across domains including mathematics [1, 27, 55, 54] and coding [8, 30, 53]. Through reinforcement learning (RL) [72, 73, 52], these models analyze problems progressively with a comprehensive Chain-of-Thought (CoT) [84, 36, 24, 32, 105, 23] before providing answers, significantly enhancing output accuracy.

The CoT reasoning strategies have also been extended to the visual domain. Recent Large Multi-modal Models (LMMs) [7, 59, 103, 92] have adapted the paradigm to accommodate the visual understanding task [51, 105, 32, 22]. These advanced LMMs can jointly process images and their associated textual queries, performing step-by-step analyses of visual details and integrating them with reasoning steps to derive final answers. Concurrently, CoT-like reasoning has been initially investigated in the visual generation task [70, 99, 74, 37, 35, 10, 34, 100], particularly in autoregressive text-to-image generation. The pioneering work, *'Image Generation with CoT'* [24], regards the progressive generation of the image tokens as a kind of CoT analogous to that of the text tokens, and proposes to optimize this intermediate process to enhance the image quality.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

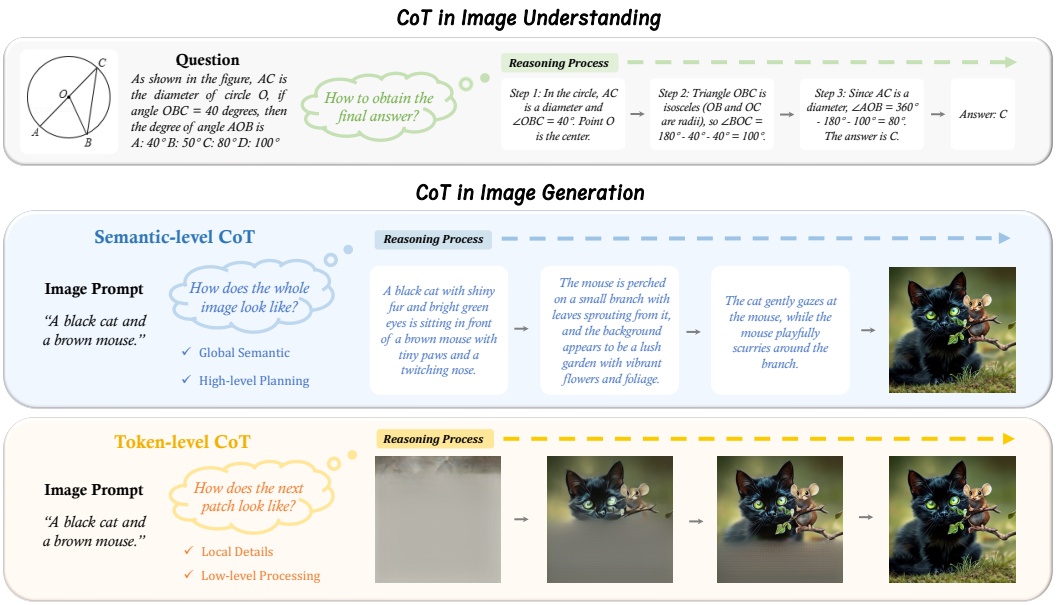

Figure 1: **The Illustration of CoT in Image Understand and Generation Tasks.** In the image understanding task, the CoT is the textual reasoning process. In the autoregressive visual generation task, we identify two levels of CoT: the semantic-level and token-level CoT. The semantic-level CoT is the high-level planning prior to the image generation, in the form of text. The token-level CoT is the intermediate patch-by-patch generation process, focusing on the local pixel details within a patch, in the form of image tokens.

Despite these advances, the exploration of CoT for image generation remains preliminary. Unlike image understanding, image generation requires the complex interpretation of cross-modal alignment and the synthesis of fine-grained visual details. To address these challenges, we identify two distinct levels of CoT reasoning that can be leveraged to enhance image generation, as illustrated in Fig. 1:

- **Semantic-level CoT** is the textual reasoning about the image to generate, which is introduced prior to the image generation. The semantic-level CoT designs the global structure of the image, e.g., the appearance and location of objects. In case the prompt requires reasoning shown in Fig. 2, the semantic-level CoT also helps to deduce the objects to generate. Optimizing the semantic-level CoT could explicitly decouple the planning and reasoning of the prompt from the subsequent image tokens generation, making the generation easier.
- **Token-level CoT** is the intermediate patch-by-patch generation process of the image, as originally introduced in [24]. This process could be viewed as a form of CoT as it outputs each subsequent token conditioned on all previous tokens within a discrete space, similar to the textual CoT. Unlike semantic-level CoT, token-level CoT focuses on low-level details like pixel generation and maintaining visual coherence between adjacent patches. Optimizing the token-level CoT can enhance both the generation quality and the alignment between the prompt and the resulting images.

Despite recognizing these two levels of CoT, a critical question remains unaddressed: *How can we enhance and coordinate them for text-to-image generation?* Current mainstream generative models [76, 79, 70, 37] are trained exclusively on generation targets, lacking the explicit textual understanding required for semantic-level CoT reasoning. Although introducing a separate model (e.g., an LLM) specifically for prompt interpretation [13] is technically feasible, this approach would significantly increase computational costs, complexity, and deployment challenges. Recently, a trend has arisen to merge visual understanding and generation within a single model. Building upon LMMs, these unified LMMs (ULMs) [86, 93, 107, 9] could not only understand the visual inputs but also generate images from text prompts. However, their two capabilities are still decoupled, typically pre-trained in two independent stages, with no clear evidence that the understanding capabilities can

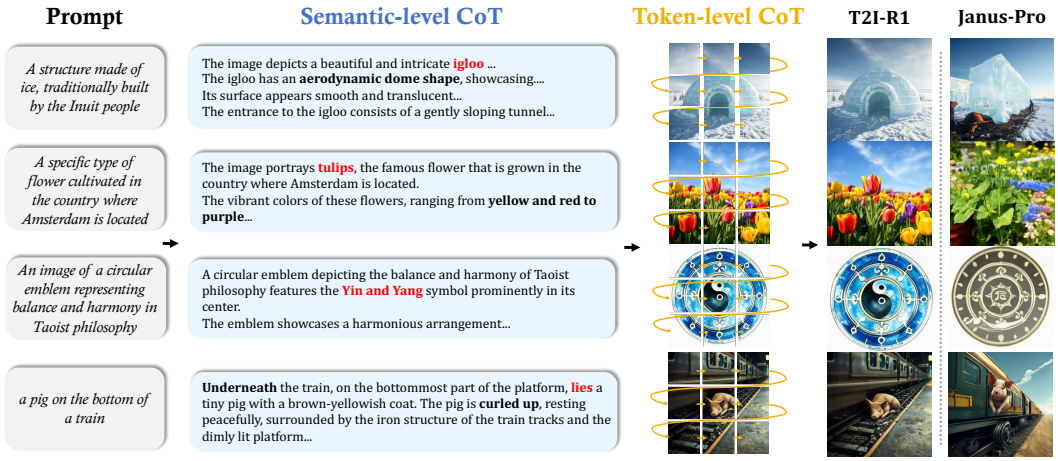

Figure 2: **Visualization of the Image Generation Process of T2I-R1.** All the prompts need reasoning or contain an uncommon scenario. We observe that T2I-R1 successfully deduces the true intention behind the prompt or provides a sensible imagination (highlighted in the text) to produce a satisfying result compared with the baseline model, Janus-Pro.

benefit generation. Given these potentials and issues, we start from a ULM and enhance it to unite both the semantic-level and token-level CoT into one framework for text-to-image generation.

To fulfill our target, we introduce **BiCoT-GRPO**, an RL method to jointly optimize the two levels of CoT for ULM. We opt for RL instead of supervised fine-tuning (SFT) for two reasons: First, the ULM has possessed the fundamental ability needed for the semantic-level and token-level CoT; our goal is only to elicit the fusion of these two abilities by guiding the model's self-exploration. Second, RL methods have proven highly effective for enhancing reasoning capabilities, which are essential for both levels of CoT. Specifically, we first instruct the ULM to imagine and plan the image based on the prompt to obtain the semantic-level CoT. Then, we feed it into the ULM as the condition for the subsequent image generation for token-level CoT. We simultaneously generate multiple images from each prompt and then compute group-relative reward to optimize both levels of CoT within the same iteration. Unlike understanding tasks, where clearly defined rules for rewards exist, image generation lacks such standardized rules. Therefore, we propose to utilize an ensemble of diverse vision experts [90, 82, 49, 24] as reward models. This reward design serves two critical purposes: it evaluates generated images from multiple dimensions to ensure reliable quality assessment, while also functioning as a regularization method to prevent the ULM from hacking a single reward model.

Through the proposed reasoning strategies, we obtain **T2I-R1**, the first reasoning-enhanced text-to-image model combining the semantic-level and token-level CoT. Empirical results show that our approach outperforms baseline models by 13% and 19% improvements on the T2I-CompBench and WISE benchmark, and even surpasses the previous state-of-the-art model FLUX.1. Qualitative analysis reveals that our method empowers the model to generate more human-aligned results by reasoning about the true intentions behind the prompt and demonstrates enhanced robustness when dealing with uncommon scenarios.

Our contributions are summarized as follows:

1. We identify a dual-level reasoning process in the autoregressive image generation task by introducing the semantic-level and token-level CoT, which decouple high-level image planning from low-level pixel generation for more reliable generation.
2. We develop BiCoT-GRPO, a new reinforcement learning framework that jointly optimizes both levels of CoT reasoning, seamlessly integrating the understanding capabilities of ULMs for image generation. For reward modeling, we investigate a robust reward system utilizing an ensemble of vision experts.
3. Our resulting model, T2I-R1, incorporates both levels of CoT using BiCoT-GRPO and demonstrates significant quantitative and qualitative improvements, surpassing FLUX.1 across multiple established benchmarks.

2 Method

2.1 Preliminary

Recently, the employment of reinforcement learning has been the dominant approach to elicit the reasoning capability of the large models. [73] introduces GRPO, enhancing PPO by eliminating the value function and estimating the advantage in a group-relative manner. For a specific prompt-answer pair (p, a), a group of G individual responses $\{o_i\}_{i=1}^{G}$ is sampled from the old policy $\pi_{\theta_{\text{old}}}$. Each response is then input to a reward function to obtain the individual reward $\mathcal{R}_i$. Then, the advantage of the i-th response is calculated by normalizing the rewards $\{\mathcal{R}_i\}_{i=1}^{G}$ of the group:

$$A_i = \frac{\mathcal{R}_i - \text{mean}(\{\mathcal{R}_i\}_{i=1}^{G})}{\text{std}(\{\mathcal{R}_i\}_{i=1}^{G})}. \tag{1}$$

GRPO adopts a clipped objective similar to PPO. Besides, a KL penalty term between the current policy π_θ and the reference model $\pi_{\theta_{\text{ref}}}$ is directly added in the loss function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \left(\min\left(r_{i,t}(\theta)\hat{A}_i, \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\Big)\hat{A}_i \right) - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}) \right) \right],$$

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

$$r_{i,j}(\theta) = \frac{\pi_\theta(o_{i,j} \mid q, o_{i,<j})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<j})}. \tag{2}$$

In text reasoning tasks like mathematical problem solving, the model is instructed to follow the pre-defined template to output the reasoning process and final answer. The reward functions are rule-based rewards that only check the correctness of the final answer and the output format.

2.2 Semantic-level and Token-level CoT

In the autoregressive text generation tasks of LLMs and LMMs, CoT occurs in the textual reasoning format. However, in autoregressive image generation tasks, we identify two distinct types of CoT that could enhance the image generation at different abstraction levels:

Semantic-level CoT. Semantic-level CoT is defined as the textual reasoning that precedes image generation, serving as an overall semantic planning stage for the intended image. This process mirrors human artistic creation: when given a brief prompt, an artist first thinks about the scene construction, considering object attributes, spatial relationships, and interactions. In addition to the planning for common prompts, we also observe the semantic-level CoT benefits two other scenarios. If the prompt does not directly depict the object to generate, the semantic-level CoT can reason about the true intention from the user's prompt, providing more aligned images. As illustrated in Fig. 2, the semantic-level CoT reasons that the flower cultivated in the country where Amsterdam is located is tulip. Without this semantic-level CoT, Janus-Pro fails to provide valid results. Additionally, the semantic-level CoT demonstrates importance when handling unusual or potentially ambiguous scenes. In the bottom example of Fig. 2, when given the prompt *'A pig on the bottom of a train'*, semantic-level CoT introduces the action 'lying' for the pig, creating a more sensible scenario. In contrast, direct generation without this interpretive imagination creates significant confusion for Janus-Pro. Formally, each semantic-level CoT s_i is composed of $|s_i|$ text tokens $\{s_{i,1}, s_{i,2}, ..., s_{i,|s_i|}\}$.

Token-level CoT. Unique to the image generation task, a token-level step-by-step thinking exists in the image generation process. The generation of image tokens much resembles a chain of thought: the image tokens are generated patch by patch, where the current patch is generated based on the previous ones. We define the sequential generation of image tokens as token-level CoT. This process parallels how an artist progressively fills a canvas, with the generated patches forming a visual reasoning chain. The reasoning content is the choice of the specific visual token for each patch, which corresponds to the patch coherence, object appearance, lighting conditions, and other visual details. Note that,

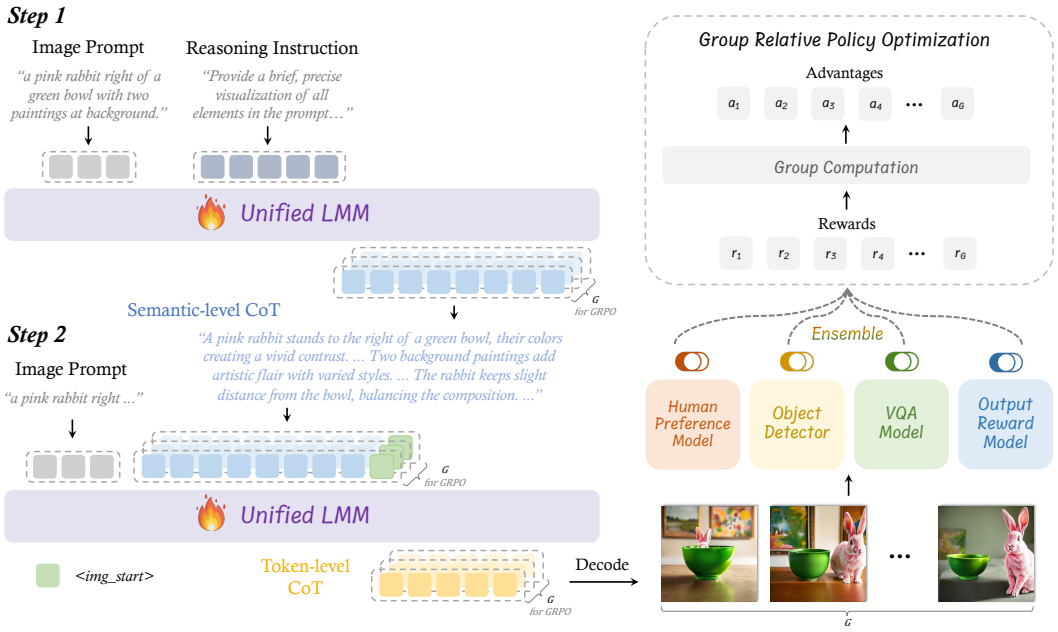

Figure 3: **Framework of BiCoT-GRPO.** In step 1, we instruct the model to generate the semantic-level CoT based on the image prompt. In step 2, images are generated conditioned on both the image prompt and semantic-level CoT, with the intermediate generation process serving as token-level CoT. The resulting images are evaluated by an ensemble of vision experts to obtain rewards. We generate N images from each prompt to compute the group-relative reward and perform GRPO training.

the reasoning occurs simultaneously with final output generation, rather than as separate steps. The model reasons and generates in parallel, integrating its thinking directly into the creation process. The generated chain of patches is later reshaped to a 2D grid $G \in \mathbb{R}^{h \times w \times c}$ and input to an image decoder $\mathbb{D}$ to obtain the image. Unlike semantic-level CoT, which addresses global planning, token-level CoT focuses on local details and visual coherence across the image space. Formally, each token-level CoT t_i consists of M image tokens $\{t_{i,1}, t_{i,2}, ..., t_{i,M}\}$, where M represents the resolution of the generated image, i.e., $M = h \times w$.

2.3 BiCoT-GRPO

GRPO has been proven to be highly effective for exploring the reasoning capability of the LLMs and LMMs. To accommodate both semantic-level and token-level CoT in image generation, we propose BiCoT-GRPO, where the model reasons twice in a single generation process. We instruct the model to first perform semantic-level CoT for global planning, and then dive into the local details by performing token-level CoT.

However, compared with the task of text generation, a great pipeline challenge is posed for incorporating two levels of CoT for image generation. Limited by the training paradigm, most current ULMs cannot generate interleaved images and text themselves. A manual signifier is often needed to instruct the model on which task to perform, either text generation or image generation. For Janus-Pro to generate an image, which is the ULM we use in this work, we need to manually concatenate an image start token (`<img_start>`) to explicitly instruct the model to start generating image tokens.

To tackle this problem, we propose a novel pipeline to facilitate ULM in generating images with two levels of CoT, as shown in Fig. 3. Specifically, our pipeline is composed of a two-step generation process. The first step is to generate the semantic-level CoT. We input the image prompt and instruct the model to imagine and reason about the details of the image to generate semantic-level CoT $\{s_i\}_{i=1}^{G}$. The second stage focuses on the token-level CoT generation. We input the image prompt, the generated semantic-level CoT in the first stage, and the image start token to the ULM for generating image tokens $\{t_i\}_{i=1}^{G}$. Then, the image tokens are input to the image decoder to obtain the

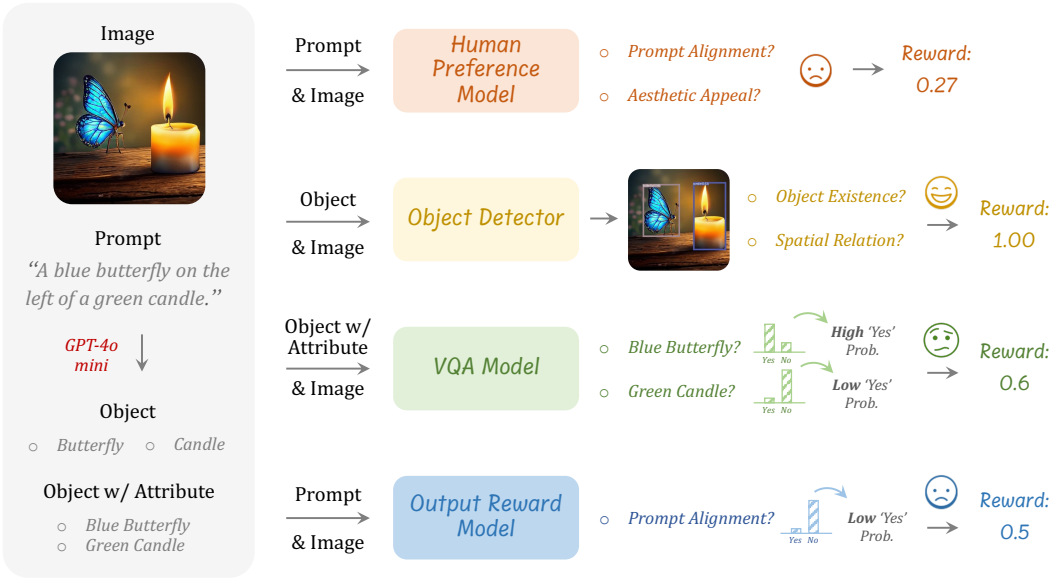

Figure 4: **Illustration of the Ensemble of Generation Rewards.** We use GPT-4o mini to extract the objects and their attributes before training. Each specialized reward model receives customized information inputs for the reward calculation. We take the average of all the rewards as final reward.

image I. Since there exist two types of CoT in our method, first the semantic-level CoT and then the token-level CoT. Each response o_i is composed of two parts, namely $o_i = (s_i, t_i)$. In this sense, the $r_{i,j}(\theta)$ is converted to:

$$r_{i,j}(\theta) = \frac{\pi_\theta(o_{i,j} \mid q, o_{i,<j})}{\pi_{\theta_{\text{old}}}(o_{i,j} \mid q, o_{i,<j})} = \begin{cases} \frac{\pi_\theta(s_{i,j}|q,s_{i,<j})}{\pi_{\theta_{\text{old}}}(s_{i,j}|q,s_{i,<j})}, & 0 \le j \le |s_i| \\ \frac{\pi_\theta(t_{i,j}|q,s_i,t_{i,<j})}{\pi_{\theta_{\text{old}}}(t_{i,j}|q,s_i,t_{i,<j})}, & |s_i| < j \le |s_i| + M \end{cases} \tag{3}$$

Then, we update the ULM by maximizing Equation 2.1. In practice, we incorporate the token-level policy gradient loss in [101], where the loss term is normalized over all the generated tokens to balance the reward on overly long semantic-level CoT.

2.4 Ensemble of Generation Rewards

Unlike DeepSeek-R1 with the rule-based reward, assessing the images based on pre-defined rules is infeasible. The assessment of the image includes various aspects, including the aesthetic appeal and objects' existence, attributes, and relationships. Considering the complexity, we introduce an ensemble of vision experts to judge the generated image from multiple aspects. Meanwhile, the use of multiple reward functions also serves as a regularization method to prevent the ULM from hacking into a specific reward model. As shown in Fig. 4, the ensemble contains the following experts:

Human Preference Model. Human preference models (HPMs), such as HPS [90] and ImageReward [94], are trained to simulate human aesthetic preferences. These models are developed using datasets of human rankings on synthetic images, where annotators evaluate and compare generated outputs. During inference, these models assess both the aesthetic quality and prompt alignment of a generated image, producing a composite human preference score $\mathcal{R}_{\text{HPM}}$. This expert provides a holistic reward signal from a general perspective.

Object Detector. Another option of the reward model is an object detector, e.g., GroundingDINO [49] and YOLO-world [12]. These open-vocabulary detection models accept an image along with object queries as input and output both the spatial positions and confidence scores for detected objects. This kind of vision expert serves as an ideal tool to evaluate the object's existence

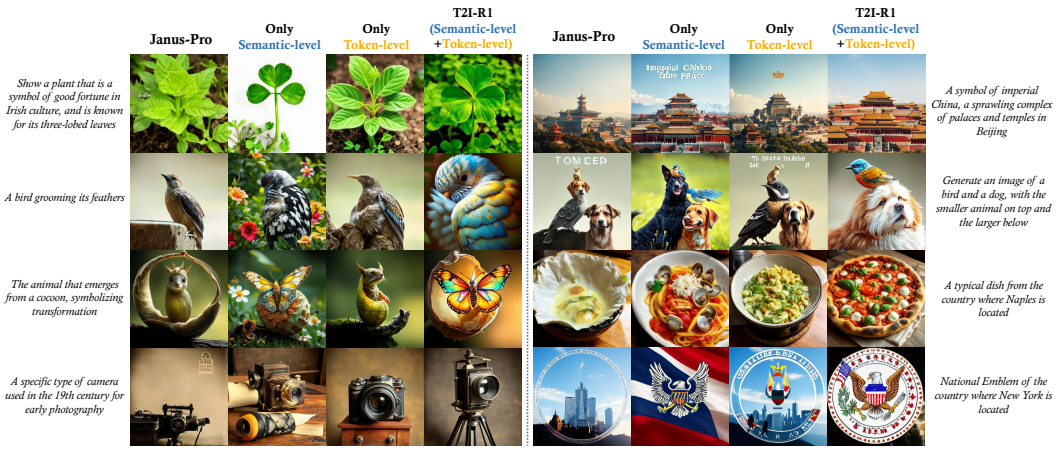

Figure 5: **Visualization Results.** We provide the image generation results of the same prompt from four models: base model, the model with only semantic-level CoT optimized, the model with only token-level CoT optimized, and the model with both levels of CoT optimized.

and relationship concerning space and numbers. For implementation, we extract all objects $\{obj_i\}_{i=1}^{K}$ from the training image prompts, where K represents the total number of objects. We then query the object detector to identify these objects within the generated image. For each object, we assign a binary existence score (1 if detected, 0 otherwise) and average these scores across all objects in the prompt. If the prompt contains a spatial relationship, we further leverage the detected location to validate its correctness. We calculate the relative distance and intersection over union (IoU) between the objects for the spatial score $\mathcal{R}_{\text{spatial}}$. If the number of the object n_{obj_i} is specifically pointed out in the prompt, we compare the number with the detected number of the object $\hat{n}_{obj_i}$. The reward from the object detector $\mathcal{R}_{\text{Det}}$ is determined as:

$$\mathcal{R}_{\text{Det}} = \begin{cases} \alpha\mathcal{R}_{\text{spatial}} + (1-\alpha)\frac{1}{K}\sum_{i=1}^{K}\mathbb{I}(obj_i \text{ detected}), & \text{if spatial relationship in the prompt,} \\ \frac{1}{n}\sum_{i=1}^{K}\mathbb{I}(n_{obj_i} = \hat{n}_{obj_i}), & \text{if number in the prompt,} \\ \frac{1}{n}\sum_{i=1}^{K}\mathbb{I}(obj_i \text{ detected}), & \text{else,} \end{cases}$$

where $\mathcal{R}_{\text{spatial}}$ is 1 if the relative distance between the objects is larger than a threshold and the direction is right. If the direction is wrong, the reward is 0. Otherwise, we use the IoU as the spatial reward. We set α as 0.6 to encourage the correctness of the spatial relationship.

Visual Question Answering Model. The visual question answering (VQA) models are trained to answer questions based on the image input. The VQA models include earlier models prior to LLM, e.g., BLIP [42] and GIT [82], and LMMs like LLaVA [47]. We leverage these models to judge the existence and attributes of the objects. For example, if the image prompt is *a red dog and a yellow cat*, we first reformat each individual object with its attribute as a question to the VQA model, i.e., *a red dog?* and *a yellow cat?*. Then, we record the probability for the model to answer *Yes* as P^i_{Yes} and *No* as P^i_{No}. The reward for a prompt is calculated as: $\mathcal{R}_{\text{VQA}} = \frac{1}{K}\sum_i \frac{P^i_{\text{Yes}}}{P^i_{\text{Yes}}+P^i_{\text{No}}}$.

Output Reward Model. Lastly, we also employ the output reward model (ORM) proposed in [24] as a reward model. The ORM is fine-tuned from an LMM (e.g., LLaVA-OneVision [39]) specifically for evaluating the alignment between the prompt and the image. The fine-tuning is to instruct the model to output *Yes* if the image perfectly aligns with the image and *No* otherwise. We calculate $\mathcal{R}_{\text{ORM}}$ using the methodology similar to $\mathcal{R}_{\text{VQA}}$, except that we input the whole image prompt to the ORM instead of reformatting the prompt. The major difference between the ORM and HPMs is that the ORM model incorporates extensive world knowledge inside the LMM while HPMs mainly focus on the human preferences including prompt-image alignment and aesthetic appeal.

We can choose one or multiple reward functions illustrated above, and take the average as the final reward for a specific sample. The detailed experiments of reward model are shown in Table 3.

Table 1: **T2I-CompBench Result.** The best score is in blue, with the second-best score in green.

Model	Attribute Binding			Object Relationship		Complex↑
	Color ↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	
Diffusion Models						
SD-v1.5 [70]	0.3758	0.3713	0.4186	0.1165	0.3112	0.3047
PixArt-α [6]	0.6690	0.4927	0.6477	0.2064	0.3197	0.3433
CoMat [31]	0.7827	0.5329	0.6468	0.2428	0.3187	0.3680
SD-XL-base-1.0 [66]	0.5879	0.4687	0.5299	0.2131	0.3119	0.3237
FLUX.1 [37]	0.7407	0.5718	0.6922	0.2863	0.3127	0.3703
AutoRegressive Models						
Show-o [93]	0.5623	0.4178	0.4641	0.2015	0.3067	0.2992
Show-o + PARM [24]	0.7549	0.5632	0.6684	0.2971	0.3126	0.3701
EMU3 [83]	0.7544	0.5706	0.7164	-	-	-
Janus-Pro-7B [9] (Baseline)	0.6359	0.3528	0.4936	0.2061	0.3085	0.3559
T2I-R1 (Ours)	0.8130	0.5852	0.7243	0.3378	0.3090	0.3993

3 Experiment

In this section, we first provide the main results of T2I-R1 in T2I-CompBench [28], WISE [61] and GenAI-Bench [45] in Section 3.1. Then we present the results of different reward function combinations in Section 3.2 and the ablation study of the effectiveness of two levels of CoT in Section 3.3. Please refer to the Appendix B for TIIF-Bench [85] results, detailed experiment setup, and more visualizations.

3.1 Main Results

We compare T2I-R1 with leading text-to-image diffusion and autoregressive models on the T2I-CompBench and WISE benchmarks (in Table 1, 2 and 4). We also provide the qualitative results in Fig. 5. Our method demonstrates substantial improvements over the baseline model, with average enhancements of 13% and 19% on T2I-CompBench and WISE, respectively. On T2I-CompBench, the most significant gains appear in attribute binding, with an average improvement of 19%. For the WISE benchmark, improvements are more evenly distributed across categories. When compared to the more powerful state-of-the-art diffusion models, T2I-R1 achieves superior or comparable results across both benchmarks. Notably, on T2I-CompBench, our method leads in five of six subtasks, with an exceptional performance in the spatial subtask (0.3378), surpassing previous SOTA results by over 5%. Similarly, for WISE, T2I-R1 excels in four of seven subtasks and achieves the highest overall score of 0.54, outperforming the robust FLUX.1-dev by 4%. Remarkably, our approach consistently achieves the leading results across all subtasks in both benchmarks when compared to other autoregressive models. Remarkably, the improvement on T2I-Compbench benefits from the planning ability brought by the semantic-level CoT, which designs the complex scenarios before generation. While the enhancement of WISE is due to the reasoning capability from the semantic-level CoT, which deduces the true object or place depicted behind the prompt. For GenAI-Bench, T2I-R1 largely improves the baseline model, and in the meantime, achieves the highest overall score on both the basic and advanced prompts. Again, T2I-R1 surpasses FLUX.1 [37] in both types of prompts and showcases a remarkable margin in the advanced prompt, probably attributed to the high-level reasoning capability granted by semantic-level CoT.

3.2 Reward Analysis

In this section, we experiment with the choice of reward functions and their combinations. We hope to provide some insights into how to choose the reward functions and combine them. Our results are shown in Table 3. We first experiment with the individual reward model. HPM (`H`) demonstrates superior performance in attribute binding but shows limited effectiveness in object relationships, likely due to its weak relation comprehension capabilities. The object detector (`D`) yields the least improvement in attribute binding, which aligns with expectations since our detector-based reward functions do not explicitly evaluate attributes. The improvements observed stem solely from enhanced object existence ratios in the prompts. We observe that VQA model (`V`) and ORM (`O`) are

Table 2: **WISE Result.** The best score is in blue, with the second-best score in green.

Model	Cultural↑	Spatio-Temporal		Natural Science			Overall
		Time↑	Space↑	Biology ↑	Physics↑	Chemistry↑	
Diffusion Models							
PixArt-Alpha [6]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Playground-v2.5 [40]	0.49	0.58	0.55	0.43	0.48	0.33	0.49
SD-v1-5 [70]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SD-XL-base-0.9 [66]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
FLUX.1-dev [37]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
AutoRegressive Models							
Emu3 [83]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Show-o [93]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
VILA-U [91]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Janus-Pro-7B [9] (Baseline)	0.30	0.37	0.49	0.36	0.42	0.26	0.35
T2I-R1 (Ours)	0.56	0.55	0.63	0.54	0.55	0.30	0.54

Table 3: **T2I-CompBench Results with Different Reward Models.** 'Det' stands for object detector.

Model	Reward Model				Attribute Binding			Object Relationship		Complex↑	Visual Quality↑
	HPM	Det	VQA	ORM	Color ↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑		
Janus-Pro-7B	-	-	-	-	0.6359	0.3528	0.4936	0.2061	0.3085	0.3559	-
-	✓	-	-	-	0.8134	0.6048	0.7311	0.2383	0.3012	0.3899	-
-	-	✓	-	-	0.7422	0.5140	0.6494	0.3044	0.3100	0.3872	-
-	-	-	✓	-	0.8171	0.6019	0.7307	0.2969	0.3088	0.4052	0.218
-	-	-	-	✓	0.7819	0.5638	0.7010	0.3301	0.3103	0.3959	1.775
-	✓	✓	-	-	0.8210	0.6074	0.7440	0.3189	0.3076	0.4005	1.942
T2I-R1	✓	✓	✓	-	0.8130	0.5852	0.7243	0.3378	0.3090	0.3993	2.063
-	✓	✓	✓	✓	0.7599	0.5742	0.6902	0.2796	0.3070	0.3921	-

both effective reward models with distinct strengths: the VQA model excels at improving attribute binding, while ORM demonstrates superior performance in relationships. Then we experiment with multiple reward models. We start from the composition of HPM and object detector ($\texttt{H} + \texttt{D}$), and progressively incorporate other reward models. Our findings indicate that both the HPM-object detector combination ($\texttt{H} + \texttt{D}$) and the three-model integration of HPM, object detector, and VQA ($\texttt{H} + \texttt{D} + \texttt{V}$) deliver balanced and satisfactory results in both attribute and relationship tasks. To obtain the optimal choice of reward models, we conduct a human study to evaluate the visual quality, detailed in Appendix C.2. We adopt the combination of the highest visual quality, the ensemble of three reward models ($\texttt{H} + \texttt{D} + \texttt{V}$) for our final model.

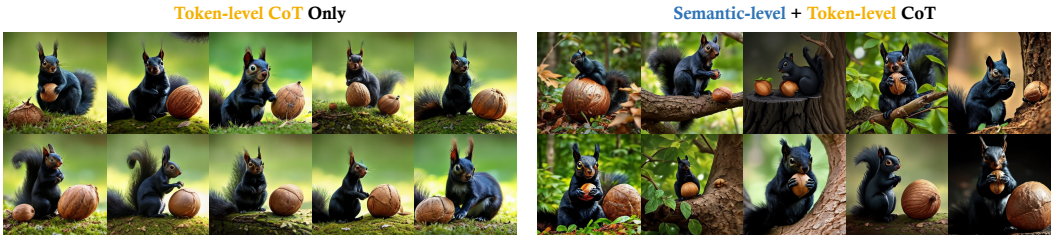

Figure 6: **Visualization Result of the Image Diversity of a Single Prompt.** We showcase the result of only token-level CoT optimized and both semantic-level and token-level CoT optimized.

3.3 Ablation Study

To validate the effectiveness of the semantic-level CoT, we compare T2I-R1 with a baseline method that generates images using only the token-level CoT optimized with the GRPO method. This is the default text-to-image generation setting in Janus, whose result is shown in the third row in Table 5. Comparing the third and fourth row in the table, we find that semantic-level CoT generally brings performance improvements across both benchmarks tested. We witness a particularly significant gain on the WISE benchmark. This enhanced performance can be attributed to the textual reasoning

Table 4: **GenAI-Bench Evaluation Results.** The best score is in blue, with the second-best score in green.

Method	Basic Prompt						Advanced Prompt					
	Attribute↑	Scene↑	Relation			Overall↑	Count↑	Differ↑	Compare↑	Logical		Overall↑
			Spatial↑	Action↑	Part↑					Negate↑	Universal↑	
Diffusion Models												
SD v2.1 [70]	0.80	0.79	0.76	0.77	0.80	0.78	0.68	0.70	0.68	0.54	0.64	0.62
SD-XL [66]	0.84	0.84	0.82	0.83	0.89	0.83	0.71	0.73	0.69	0.50	0.66	0.63
Midjourney v6 [60]	0.88	0.87	0.87	0.87	0.91	0.87	0.78	0.78	0.79	0.50	0.76	0.69
FLUX.1-dev [37]	0.87	0.88	0.87	0.85	0.87	0.87	0.75	0.78	0.74	0.45	0.70	0.64
Auto-Regressive Models												
LWM [46]	0.63	0.62	0.65	0.63	0.70	0.63	0.59	0.58	0.54	0.49	0.52	0.53
Show-o [93]	0.72	0.72	0.70	0.70	0.75	0.70	0.70	0.62	0.71	0.51	0.65	0.60
VILA-U [91]	0.78	0.78	0.77	0.78	0.79	0.76	0.70	0.71	0.74	0.53	0.66	0.64
Liquid [87]	–	–	–	–	–	–	0.76	0.73	0.74	0.46	0.74	0.65
UniTok [56]	–	–	–	–	–	–	0.76	0.76	0.79	0.46	0.73	0.67
Mogao-7B [44]	–	–	–	–	–	–	0.77	0.74	0.77	0.53	0.71	0.68
Janus-Pro-7B [9] (Baseline)	0.85	0.87	0.85	0.84	0.85	0.84	0.73	0.73	0.71	0.48	0.65	0.65
T2I-R1 (Ours)	0.87	0.89	0.89	0.87	0.87	0.88	0.81	0.82	0.78	0.60	0.73	0.73

Table 5: **Ablation Experiments on the Effectiveness of the Two Levels of CoT.**

Model	Optimized CoT		T2I-CompBench			WISE			Diversity↑
	Semantic-level	Token-level	Color↑	Shape↑	Texture↑	Culture↑	Spatio-Temporal↑	Science↑	
Janus-Pro-7B			0.6359	0.3528	0.4936	0.3000	0.4232	0.3467	6.976
-	✓		0.8082	0.5684	0.7219	0.4900	0.5599	0.4367	8.177
-		✓	0.7752	0.5849	0.7451	0.3500	0.4732	0.3900	6.255
T2I-R1	✓	✓	0.8130	0.5852	0.7243	0.5600	0.5855	0.4633	8.203

capabilities inherent in semantic-level CoT. As illustrated in Fig. 5, our method could first clearly reason about the objects or phenomena described in the prompt through semantic-level CoT. This effectively decouples the reasoning and generation processes and thereby facilitates superior results. We also observe that training solely with token-level CoT substantially reduces the diversity of generated images, as demonstrated in Fig. 6, 7, 13, and 14. To quantify this effect, we evaluate image diversity by reusing the generated images from T2I-CompBench, where each prompt generates ten images. We compute the Vendi Score [18] across the ten images for each prompt. Results indicate that GRPO training without semantic-level CoT decreases the diversity score, whereas incorporating semantic-level CoT significantly improves diversity through varied textual planning.

We also consider another situation to validate the effectiveness of token-level CoT: the semantic-level CoT is incorporated in the image generation process, as T2I-R1, but GRPO only optimizes the semantic-level CoT without the token-level CoT. This can be viewed as only enhancing the model's high-level planning capabilities. The second row of Table 5 presents the result. The results show that optimizing semantic-level CoT exclusively yields smaller improvements compared to the joint optimization approach. Additionally, we find that optimizing both CoT types produces images with much better aesthetic quality compared with optimizing semantic-level CoT only, as shown in Fig. 5. This indicates the necessity to jointly optimize both levels of CoT.

4 Conclusion

In this paper, we introduce T2I-R1, the first reasoning-enhanced text-to-image model powered by a bi-level CoT reasoning process. We identify the semantic-level CoT for high-level planning and the token-level CoT for patch-by-patch generation. We further integrate them through our proposed BiCoT-GRPO, an RL framework incorporating two levels of CoT within the same training step. By leveraging a ULM capable of both visual understanding and generation, our approach eliminates the need for separate specialized models while achieving significant performance improvements, +13% on T2I-CompBench and +19% on the WISE benchmark. Our qualitative analysis demonstrates that T2I-R1 better understands complex prompts, reasons about user intentions, and handles uncommon scenarios with greater robustness, establishing a new paradigm for text-to-image generation tasks.

Acknowledgments

This study was supported in part by National Key R&D Program of China Project 2022ZD0161100, in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, in part by NSFC-RGC Project N_CUHK498/24, and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515130008, XW).

Additional support was provided by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project 14201321 and Project 14200824.

References

[1] Amini, A., Gabriel, S., Lin, P., Koncel-Kedziorski, R., Choi, Y., Hajishirzi, H.: Mathqa: Towards interpretable math word problem solving with operation-based formalisms. arXiv preprint arXiv:1905.13319 (2019)

[2] Austin, J., Odena, A., Nye, M.I., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C.J., Terry, M., Le, Q.V., Sutton, C.: Program synthesis with large language models. CoRR **abs/2108.07732** (2021)

[3] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)

[4] Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)

[5] Chen, J., Xue, L., Shu, M., Zhang, Y., Zhang, Y., Xu, R., Wang, X.E., Xiong, C.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)

[6] Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)

[7] Chen, L., Li, L., Zhao, H., Song, Y., Vinci: R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V (2025), accessed: 2025-02-02

[8] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W.: Evaluating large language models trained on code. CoRR **abs/2107.03374** (2021)

[9] 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)

[10] Chen, X., Ge, J., Zhang, T., Liu, J., Zhang, S.: Learning from mistakes: Iterative prompt relabeling for text-to-image diffusion model training. arXiv preprint arXiv:2312.16204 (2023)

[11] Chen, X., Zhang, R., Jiang, D., Zhou, A., Yan, S., Lin, W., Li, H.: Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. arXiv preprint arXiv:2506.05331 (2025)

[12] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16911 (2024)

[13] Datta, S., Ku, A., Ramachandran, D., Anderson, P.: Prompt expansion for adaptive text-to-image generation. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3449–3476. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.189, https://aclanthology.org/2024.acl-long.189/

[14] Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement (2025), https://arxiv.org/abs/2503.17352

[15] Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)

[16] Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)

[17] Fang, R., Duan, C., Wang, K., Li, H., Tian, H., Zeng, X., Zhao, R., Dai, J., Li, H., Liu, X.: Puma: Empowering unified mllm with multi-granular visual generation. arXiv preprint arXiv:2410.13861 (2024)

[18] Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410 (2022)

[19] Gemini Team, G.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[20] 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)

[21] Guo, Z., Lin, H., Yuan, Z., Zheng, C., Qiu, P., Jiang, D., Zhang, R., Feng, C.M., Li, Z.: Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. arXiv preprint arXiv:2503.10529 (2025)

[22] Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., Heng, P.A.: Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. arXiv preprint arXiv:2510.26802 (2025)

[23] Guo, Z., Zhang, R., Chen, H., Gao, J., Jiang, D., Wang, J., Heng, P.A.: Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. arXiv preprint arXiv:2503.10627 (2025)

[24] Guo, Z., Zhang, R., Tong, C., Zhao, Z., Gao, P., Li, H., Heng, P.A.: Can we generate images with cot? let's verify and reinforce image generation step by step. arXiv preprint arXiv:2501.13926 (2025)

[25] Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis (2024)

[26] He, J., Li, H., Hu, Y., Shen, G., Cai, Y., Qiu, W., Chen, Y.C.: Disenvisioner: Disentangled and enriched visual prompt for customized image generation. arXiv preprint arXiv:2410.02067 (2024)

[27] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. NeurIPS (2021)

[28] Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)

[29] Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)

[30] Jain, N., Han, K., Gu, A., Li, W., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., Stoica, I.: Livecodebench: Holistic and contamination free evaluation of large language models for code. CoRR **abs/2403.07974** (2024), https://doi.org/10.48550/arXiv.2403.07974

[31] Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., Li, H.: Comat: Aligning text-to-image diffusion model with image-to-text concept matching. arXiv preprint arXiv:2404.03653 (2024)

[32] Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al.: Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621 (2025)

[33] Jiang, D., Zhang, R., Guo, Z., Wu, Y., Lei, J., Qiu, P., Lu, P., Chen, Z., Song, G., Gao, P., et al.: Mmsearch: Benchmarking the potential of large models as multi-modal search engines. arXiv preprint arXiv:2409.12959 (2024)

[34] Jiang, D., Zhang, R., Li, H., Zong, Z., Guo, Z., He, J., Guo, C., Ye, J., Fang, R., Li, W., et al.: Draco: Draft as cot for text-to-image preview and rare concept generation. arXiv preprint arXiv:2512.05112 (2025)

[35] Jiang, D., Zhang, R., Shu, Y., Peng, Y., Zong, Z., Duan, Y., Wang, Z., Liu, J., Chen, H., Guo, Z., Ye, J., Liu, R., Heng, P.A., Zhang, S., Li, H.: Ulmevalkit: An open-source toolkit for evaluating unified large multi-modal models and generative models (October 2025), https://github.com/ULMEvalKit/ULMEvalKit

[36] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

[37] Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)

[38] Lei, J., Zhang, R., Hu, X., Lin, W., Li, Z., Sun, W., Du, R., Zhuo, L., Li, Z., Li, X., et al.: Imagine-e: Image generation intelligence evaluation of state-of-the-art text-to-image models. arXiv preprint arXiv:2501.13920 (2025)

[39] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)

[40] Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024)

[41] Li, H., Tian, C., Shao, J., Zhu, X., Wang, Z., Zhu, J., Dou, W., Wang, X., Li, H., Lu, L., et al.: Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. arXiv preprint arXiv:2412.09604 (2024)

[42] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

[43] Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. arXiv: 2403.18814 (2024)

[44] Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025)

[45] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)

[46] Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with ringattention. arXiv e-prints pp. arXiv–2402 (2024)

[47] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

[48] Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al.: Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. arXiv preprint arXiv:2503.10631 (2025)

[49] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., yue Li, C., Yang, J., Su, H., Zhu, J.J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. ArXiv **abs/2303.05499** (2023)

[50] Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)

[51] Lu, P., Bansal, H., Xia, T., Liu, J., yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. ArXiv **abs/2310.02255** (2023)

[52] Lu, Z., Ren, H., Yang, Y., Wang, K., Zong, Z., Pan, J., Zhan, M., Li, H.: Webgen-agent: Enhancing interactive website generation with multi-level feedback and step-level reinforcement learning (2025), https://arxiv.org/abs/2509.22644

[53] Lu, Z., Yang, Y., Ren, H., Hou, H., Xiao, H., Wang, K., Shi, W., Zhou, A., Zhan, M., Li, H.: Webgen-bench: Evaluating llms on generating interactive and functional websites from scratch (2025), https://arxiv.org/abs/2505.03733

[54] Lu, Z., Zhou, A., Ren, H., Wang, K., Shi, W., Pan, J., Zhan, M., Li, H.: Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms (2024), https://arxiv.org/abs/2402.16352

[55] Lu, Z., Zhou, A., Wang, K., Ren, H., Shi, W., Pan, J., Zhan, M., Li, H.: Mathcoder2: Better math reasoning from continued pretraining on model-translated mathematical code (2024), https://arxiv.org/abs/2410.08196

[56] Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025)

[57] Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Zhao, L., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. arXiv preprint arXiv:2411.07975 (2024)

[58] MAA: American invitational mathematics examination - aime. In: American Invitational Mathematics Examination - AIME 2024 (February 2024), https://maa.org/math-competitions/american-invitational-mathematics-examination-aime

[59] Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Shi, B., Wang, W., He, J., Zhang, K., et al.: Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)

[60] Midjourney: Midjourney v6.1. https://www.midjourney.com/ (2024)

[61] Niu, Y., Ning, M., Zheng, M., Lin, B., Jin, P., Liao, J., Ning, K., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)

[62] OpenAI: Chatgpt. https://chat.openai.com (2023)

[63] OpenAI: GPT-4V(ision) system card (2023), https://openai.com/research/gpt-4v-system-card

[64] OpenAI: Hello gpt-4o. https://openai.com/index/hello-gpt-4o/ (2024)

[65] OpenAI: Introducing openai o1, 2024. (2024), https://openai.com/o1/

[66] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[67] Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. arXiv preprint arXiv:2412.03069 (2024)

[68] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445

[69] Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J., Michael, J., Bowman, S.R.: GPQA: A graduate-level google-proof q&a benchmark. arXiv preprint arXiv:2311.12022 (2023)

[70] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[71] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

[72] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[73] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)

[74] Shen, D., Song, G., Zhang, Y., Ma, B., Li, L., Jiang, D., Zong, Z., Liu, Y.: Adt: Tuning diffusion models with adversarial supervision. arXiv preprint arXiv:2504.11423 (2025)

[75] Song, W., Wang, Y., Song, Z., Li, Y., Sun, H., Chen, W., Zhou, Z., Xu, J., Wang, J., Yu, K.: Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. arXiv preprint arXiv:2503.14324 (2025)

[76] Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)

[77] Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. arXiv: 2312.13286 (2023)

[78] Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative pretraining in multimodality. arXiv: 2307.05222 (2023)

[79] Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)

[80] Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164 (2024)

[81] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[82] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)

[83] Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)

[84] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[85] Wei, X., Zhang, J., Wang, Z., Wei, H., Guo, Z., Zhang, L.: Tiif-bench: How does your t2i model follow your instructions? arXiv preprint arXiv:2506.02161 (2025)

[86] Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848 (2024)

[87] Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., Bai, X.: Liquid: Language models are scalable multi-modal generators. arXiv preprint arXiv:2412.04332 (2024)

[88] Wu, W., Li, Z., He, Y., Shou, M.Z., Shen, C., Cheng, L., Li, Y., Gao, T., Zhang, D., Wang, Z.: Paragraph-to-image generation with information-enriched diffusion model. International Journal of Computer Vision **133**, 5413–5434 (2025). https://doi.org/10.1007/s11263-025-02435-1

[89] Wu, X., Bai, Y., Zheng, H., Chen, H.H., Liu, Y., Wang, Z., Ma, X., Shu, W.J., Wu, X., Yang, H., Lim, S.N.: LightGen: Efficient Image Generation through Knowledge Distillation and Direct Preference Optimization (Mar 2025). https://doi.org/10.48550/arXiv.2503.08619

[90] Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)

[91] Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429 (2024)

[92] Xiao, H., Wang, G., Chai, Y., Lu, Z., Lin, W., He, H., Fan, L., Bian, L., Hu, R., Liu, L., et al.: Ui-genie: A self-improving approach for iteratively boosting mllm-based mobile gui agents. arXiv preprint arXiv:2505.21496 (2025)

[93] 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)

[94] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 15903–15935 (2023)

[95] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Fan, Z.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)

[96] Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: International Conference on Machine Learning (2024)

[97] Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)

[98] Ye, J., Jiang, D., He, J., Zhou, B., Huang, Z., Yan, Z., Li, H., He, C., Li, W.: Blink-twice: You see, but do you observe? a reasoning benchmark on visual perception. arXiv preprint arXiv:2510.09361 (2025)

[99] 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)

[100] Ye, J., Zhu, L., Guo, Y., Jiang, D., Huang, Z., Zhang, Y., Yan, Z., Fu, H., He, C., Li, W.: Realgen: Photorealistic text-to-image generation via detector-guided rewards. arXiv preprint arXiv:2512.00473 (2025)

[101] Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)

[102] Zhan, J., Ai, Q., Liu, Y., Pan, Y., Yao, T., Mao, J., Ma, S., Mei, T.: Prompt refinement with image pivot for text-to-image generation. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 941–954. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.53, https://aclanthology.org/2024.acl-long.53/

[103] Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937 (2025)

[104] Zhang, R., Han, J., Liu, C., Zhou, A., Lu, P., Qiao, Y., Li, H., Gao, P.: Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In: ICLR 2024 (2024)

[105] Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Gao, P., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? ECCV 2024 (2024)

[106] Zhang, R., Wei, X., Jiang, D., Zhang, Y., Guo, Z., Tong, C., Liu, J., Zhou, A., Wei, B., Zhang, S., et al.: Mavis: Mathematical visual instruction tuning. arXiv preprint arXiv:2407.08739 (2024)

[107] Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)

[108] Zong, Z., Jiang, D., Ma, B., Song, G., Shao, H., Shen, D., Liu, Y., Li, H.: Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. arXiv preprint arXiv:2412.09618 (2024)

[109] Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y.: Mova: Adapting mixture of vision experts to multimodal context. arXiv preprint arXiv:2404.13046 (2024)

A Related Work

Unified Generation and Understanding LMM. Recently, the effort to unify image generation and understanding in a single LMM has attracted much attention. Building upon large language models (LLMs), it is natural for the LMMs to understand the image and output the text [63, 39, 109, 19, 104, 33]. However, the method of how to generate an image from a LMM is still under exploration. The image generation method diverges into different branches. One line of the method relies on an exterior image generation model to complete generation [15, 78, 77, 43, 80, 17, 108, 38]. The generator often utilizes text-to-image diffusion models [70, 66] due to its powerful generation capability. To deliver the generation information, the LMM passes either the implicit conditional feature or the explicit image prompt to the generator. For example, EMU [78] first trains the LMM to output CLIP [68] image features identical to that input to the LMM. Then, a pretrained UNet [71] of Stable Diffusion [70] receives the output feature as the condition to generate an image. Another line of the method seeks to train the LMM to generate discrete tokens produced by VQGAN [16] to eliminate the need for an additional generator. [83, 41] directly adopts the VQGAN encoder as the image tokenizer for LMM. However, the VQGAN encoder is only pretrained on the image reconstruction task and thereby generates visual tokens less helpful for image understanding. To improve the understanding capability, [86, 9, 57, 48] proposes to tackle the understanding and generation tasks with different vision encoders separately. The CLIP encoder deals with image input for understanding, while the VQGAN encoder is responsible for generation. Moreover, some works [91, 67, 75] attempt to empower the vision encoder with both the understanding and the generation capability. VILA-U [91] trains a vision encoder with both the contrastive loss [68] for text-image understanding and reconstruction loss [16] for image detail preserving. Thanks to the joint pretraining, the vision encoder could generate text-aligned discrete visual tokens. The LMM is then trained to receive the discrete tokens for image understanding and predict them for image generation.

Reinforcement Learning for Large Reasoning Models. The emergence of OpenAI o1 [65] has gained tremendous attention in developing the reasoning capability of large language models. Later, DeepSeek-R1 [20] proposes a rule-based reward and GRPO training method. The introduced method instructs the model to perform an extensive reasoning process before generating the final answer. The reward only focuses on the correctness of the final answer and the following of the pre-defined format. Recently, a number of works have applied this method to multi-modal large language models [7, 59, 97, 103, 14, 29, 11] with task-specific rewards like correctness and IoU [50]. This training paradigm largely helps various reasoning-intensive tasks [69, 32, 23] like mathematical problem-solving [27, 58, 51, 105, 106], code generation [8, 2, 30], and complex scene understanding [98].

B More Experiment Details

B.1 Experiment Setup

Training Settings. Our training dataset comprises text prompts sourced from the training set of T2I-CompBench [28] and [24], totaling 6,786 prompts with no images. Prior to training, we use GPT-4o mini to extract the objects and their attributes from the prompts to facilitate computing the rewards. We use Janus-Pro-7B as the base model. We use a learning rate of 1e-6 and a beta of 0.01. For the reward model, we choose HPS [90] as the human preference model, GroundingDINO [49] as the object detector, and GIT [82] as the VQA model. For the ORM, we finetune LLaVA-OneVision-7B in the same manner as [24].

Benchmark. We test on T2I-CompBench [28], WISE [61], GenAI-Bench [45], and TIIF-Bench [85] to validate the effectiveness of our method. T2I-CompBench comprises 6,000 compositional text prompts evaluating three categories (attribute binding, object relationships, and complex compositions) and six sub-categories (color binding, shape binding, texture binding, spatial relationships, non-spatial relationships, and complex compositions). WISE consists of 1,000 text prompts spanning three categories (cultural common sense, spatial-temporal reasoning, and natural science) for evaluating world knowledge of the text-to-image models. To correctly generate an image, the model needs to reason about what the exact object or scenario is depicted in the prompt. We slightly modify the reasoning instruction on the WISE benchmark for more aligned results. GenAI-

Figure 7: **More Visualization Result of the Image Diversity of a Single Prompt.** We showcase the result of only token-level CoT optimized and both semantic-level and token-level CoT optimized.

Table 6: **TIIF-Bench Testmini Subset Evaluation Results.** The best score is in blue, with the second-best score in green.

Model	Overall		Basic Following								Advanced Following												Designer	
			Avg		Attribute		Relation		Reasoning		Avg		Attribute +Relation		Attribute +Reasoning		Relation +Reasoning		Style		Text		Real World	
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long
Llamagen [76]	41.67	38.22	53.00	50.00	48.33	42.33	59.57	60.32	51.07	47.32	35.89	32.61	38.82	31.57	40.84	47.22	49.59	46.22	46.67	33.33	0.00	0.00	39.73	35.62
LightGen [89]	53.22	43.41	66.58	47.91	55.83	47.33	74.82	45.82	69.07	50.57	46.74	41.53	62.44	40.82	61.71	50.47	50.34	45.34	53.33	53.33	0.00	6.83	50.92	50.55
Show-o [93]	59.72	58.86	73.08	75.83	74.83	79.83	78.82	78.32	65.57	69.32	53.67	50.38	60.95	56.82	68.59	68.96	66.46	56.22	63.33	66.67	3.83	2.83	55.02	50.92
Infinity [25]	62.07	62.32	73.08	75.41	74.33	76.83	72.82	77.57	72.07	71.82	56.64	54.98	60.44	55.57	74.22	64.71	60.22	59.71	80.00	73.33	10.83	23.83	54.28	56.89
Janus-Pro [9]	66.50	65.02	79.33	78.25	79.33	82.33	78.32	73.32	80.32	79.07	59.71	58.82	66.07	56.20	70.46	70.84	67.22	59.97	60.00	70.00	28.83	33.83	65.84	60.25
T2I-R1 (Ours)	68.59	67.19	82.90	81.63	86.50	83.00	83.47	79.43	78.73	82.46	69.05	68.00	71.64	69.47	72.43	69.95	69.40	70.40	60.00	63.33	27.60	26.24	67.54	60.45

Bench is a benchmark containing 1,600 complex, real-world text prompts collected from professional designers, which covers a broad spectrum of compositional text-to-visual generation elements, from basic aspects like scenes, attributes, and relationships to more professional ones, including counting, comparison, differentiation, and logical reasoning. TIIF-Bench is a comprehensive benchmark for fine-grained text-to-image model evaluation, featuring 36 novel prompt combinations across six compositional dimensions and 100 real-world designer-level prompts with rich aesthetic judgment. We follow the official evaluation setting of all the benchmarks.

C More Experiment Results

C.1 More Results

We provide the experiment results on TIIF-Bench in Table 6 and more qualitative examples in Fig. 8.

Finally, we discuss the zero-shot potential of the baseline model to perform both semantic-level and token-level reasoning. Specifically, we apply the same image generation process of T2I-R1 directly to the baseline model, where the baseline model is first instructed to output the semantic-level CoT and then the token-level CoT. We term this method of generation as 'Janus-Pro w/ zero-shot semantic-level CoT' in Figure 9-12. As shown in the figure, zero-shot semantic-level CoT brings very marginal improvement, while T2I-R1 demonstrates a satisfying result. The reasons are twofold: (1) Zero-shot semantic-level CoT misses critical objects in the original prompt. As shown in Figure 12, the zero-shot semantic-level CoT misses the *bird* in the original prompt. (2) Zero-shot semantic-level CoT does not fit the model's generation ability or provide useful information for generation. Although

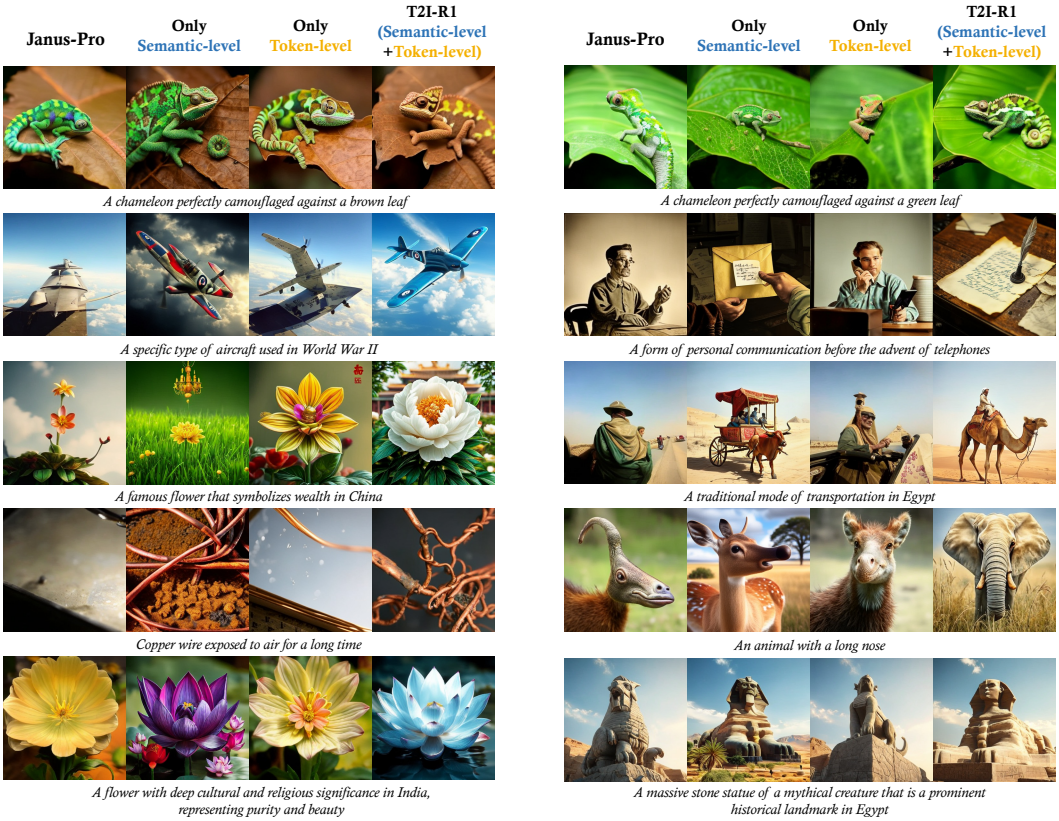

Figure 8: **More Visualization Results.** We provide the image generation results of the same prompt from four models: base model, the model with only semantic-level CoT optimized, the model with only token-level CoT optimized, and the model with both levels of CoT optimized.

the semantic-level CoT in Figure 9-11 includes all the objects and relationships, the baseline model still fails to generate a satisfying result. This highlights the necessity of our proposed BiCoT-GRPO training method to build the synergy between the two levels of CoT and make them work together.

C.2 More Details about Reward Analysis

We conduct a human study to evaluate the visual quality of the generated images. Specifically, we select four options of reward models (`V`, `O`, `H` + `D`, and `H` + `D` + `V`) to generate an image from the same prompt. Then we ask humans to rank the four images and score them according to the rank (rank 1 for 3 points, rank 2 for 2 points, and so on). We ask the annotators to judge only according to the visual appeal. We provide examples of an unsatisfactory generation, including different aspects: 1) Corrupted floating words, 2) Distortion like melting effect of faces, hands, or other objects, 3) Extra or missing hands, legs, tails, etc., 4) Object merging, frequently observed for compositional prompts, 5) Low-level aspects including over exposure or over saturation, 6) Blurred areas, especially in fine details like keyboard. The instruction to the annotator highlights that the highly ranked images should show no or fewer problems compared with others. Eight graduate students are employed to conduct the study to eliminate individual bias. We randomly choose 30 prompts from each of the subtasks from the T2I-CompBench. The result is shown in the visual quality column in Table 3. We observe that ensemble rewards achieve better visual quality, with `H` + `D` + `V` obtaining slightly superior results. This improvement could be attributed to the implicit regularization provided by multiple rewards, preventing overfitting to a single reward model. Conversely, individual reward models fail to provide satisfactory quality despite high benchmark scores.

C.3 Comparison with Prompt Rewriting Methods

The prompt rewriting [96, 102, 26, 88] is a method to leverage an external LLM to generate an enriched version of the prompt and use this version for the text-to-image model to produce the images. Our semantic-level CoT design resembles the high-level idea of this method, but there exist several key differences. First, the semantic-level CoT is generated from the image generation model itself, while the prompt enrichment leverages an extra LLM. Second, our design facilitates a joint-optimization of the prompt design and the image generation. On the contrary, the prompt enrichment is generation-model-agnostic, which means that although the prompt is enriched with more details, it is not necessary that the generation model can generate a better image based on this enriched prompt. We validate this claim in Figure 9-11. Although the zero-shot semantic-level CoT (enriched prompt) all correctly includes all the key objects mentioned in the prompt and adds more details, the model still cannot generate satisfying result. Simply enriching the prompt is not sufficient to bring notable improvements.

C.4 Comparison with Training with RL and Supervised Finetuning

In our experiments, we directly employ reinforcement learning. An alternative is to first conduct supervised finetuning (SFT) and then continue with RL training. Our findings show that a cold start stage with SFT seems to be detrimental for the final performance. Specifically, we use the open-source high-quality text-to-image dataset BLIP3o-60K [5] as our SFT dataset. We employ Qwen2.5-VL-72B-Instruct [3] to generate the semantic-level CoT with the image and the original short prompt input. We finetune the model for one epoch and then conduct RL finetuning following the same training setting of T2I-R1. The results are shown in Table 7 below:

Table 7: **Model Performance Comparison Between RL and SFT.**

Training Method	Color	Shape	Texture	Spatial	Non-Spatial	Complex
Janus-Pro	0.6359	0.3528	0.4926	0.2061	0.3085	0.3559
SFT Only	0.7035	0.5217	0.6423	0.2775	0.3068	0.3626
Hybrid optimization	0.7765	0.5832	0.6981	0.3327	0.3092	0.3949
T2I-R1 (RL Only)	0.8130	0.5852	0.7243	0.3378	0.3090	0.3993

The key findings are two-fold. First, high-quality SFT is beneficial. The curated BLIP3-60k dataset yields significant performance gains. However, SFT performance substantially lags behind RL training. We hypothesize this occurs because SFT constrains the model to replicate the training distribution rather than leveraging its inherent capabilities. Specifically, when presented with a valid semantic-level CoT, SFT forces the model to generate the exact corresponding training image, even when the model could produce alternative valid outputs. This constraint introduces unnecessary training complexity. Second, RL after high-quality SFT is still inferior to direct RL training. While high-quality SFT pre-training improves subsequent RL performance, the combined approach remains mostly inferior to or merely comparable with direct RL training.

C.5 Choice of Reward Weights

In Table 1, the weight of all the rewards are set to 1. We find that the weight of the reward model has little influence on the final training result. Here we provide the detailed study to illustrate the our method is stable to the weight of the reward model. We follow the training setting of T2I-R1, where three reward models, HPS, GroundingDINO, and GIT are employed. In our experiments, we multiply the weight of each reward model by 5 respectively, while maintaining the weight of the other two reward models. Apart from these three experiments, we also conduct an experiment where we compute the relative reward of each reward model inside the group, and then sum up the normalized reward as the final reward for the sample. This eliminates the mean and variance difference among the reward models (termed as Normalized Reward). The results are shown in Table 8 below.

C.6 Hyperparameters

All of our experiments are conducted on 8 H800. Our training procedure lasts about 16 hours. We provide the detailed training hyperparameters in Table 9.

Table 8: **Comparison of Different Reward Weights.**

Reward Design	Color	Shape	Texture	Spatial	Non-Spatial	Complex
5∗HPS, 1∗others	0.8215	0.5915	0.7337	0.3051	0.3074	0.4051
5∗GDINO, 1∗others	0.7951	0.5520	0.7100	0.3313	0.3104	0.3916
5∗GIT, 1∗others	0.7972	0.5620	0.7149	0.3357	0.3111	0.3934
Normalized Reward	0.8106	0.5820	0.7142	0.2940	0.3072	0.3993
T2I-R1 (Equal Weights)	0.8130	0.5852	0.7243	0.3378	0.3090	0.3993

Table 9: T2I-R1 training hyperparameters.

Name	
Learning rate	1e-6
Beta β	0.01
Group Size G	8
Classifier-Free Guidance Scale	5
Max Gradient Norm	1.0
Batchsize	8
Training Steps	1,600
Gradient Accumulation Steps	2
Image Resolution $h \times w$	384×384

D Limitations and Future Work

While this work explores the text-to-image generation task, it requires more exploration on how to apply this paradigm to other modalities like video generation [4] or 3D [21] tasks. Specifically, video generation tasks are more complex regarding the reward design and the base model. For the reward design, how to apply dense rewards on each generated frame is still an open question. Besides, there exists no understanding and generation unified model for videos, so BiCoT-GRPO cannot be used directly. Meanwhile, the current inference time of video generation is too long for the current GRPO paradigm. How to balance the training time and effect needs further study.

Figure 9: **Visualization Results of Semantic-level CoT.** We provide the image generation results of the same prompt from three settings: base model, base model with zero-shot semantic-level CoT, and T2I-R1. For the setting of base model with zero-shot semantic-level CoT, we use the same generation pipeline of T2I-R1 directly on the base model. We employ the same prompt of T2I-R1 to instruct the base model to generate a zero-shot semantic-level CoT, which we visualize in the figure and provide a comparison of the semantic-level CoT generated by T2I-R1.

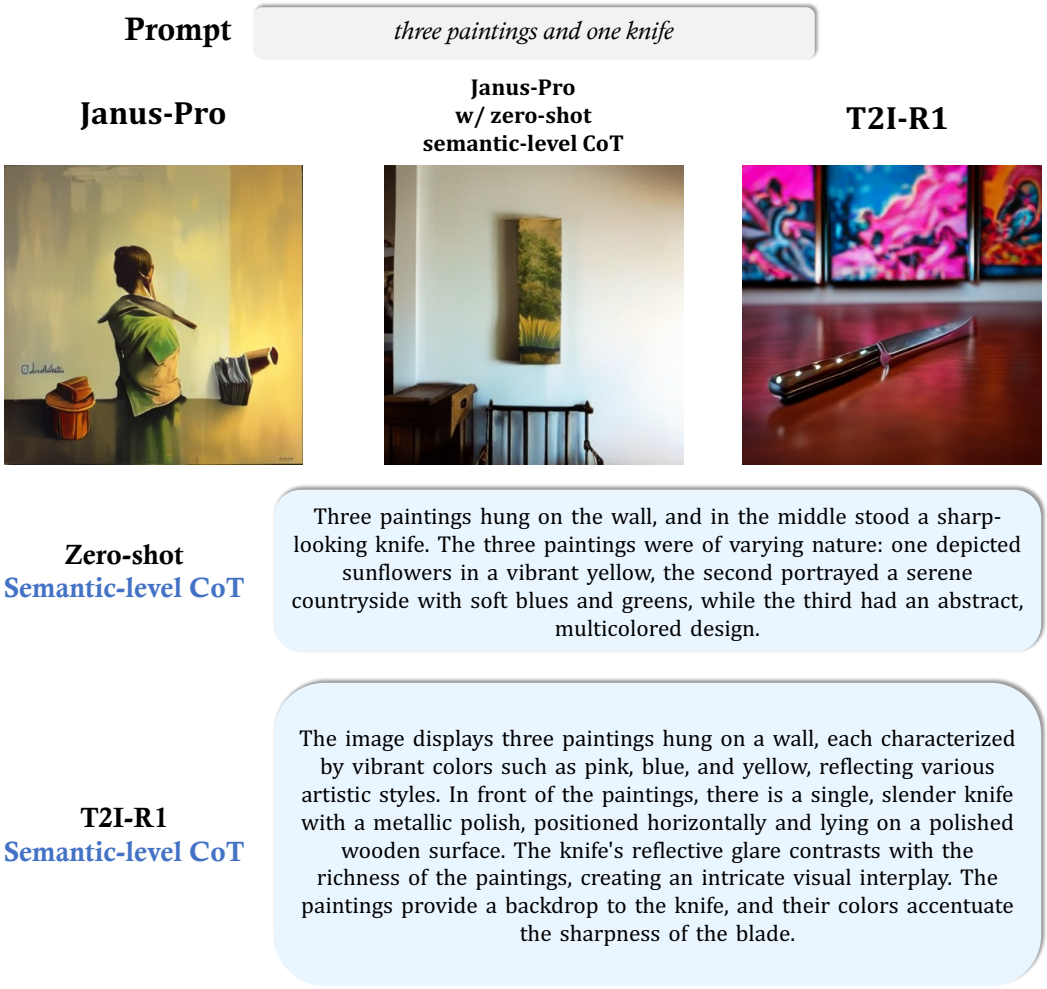

Figure 10: **Visualization Results of Semantic-level CoT.** We provide the image generation results of the same prompt from three settings: base model, base model with zero-shot semantic-level CoT, and T2I-R1. For the setting of base model with zero-shot semantic-level CoT, we use the same generation pipeline of T2I-R1 directly on the base model. We employ the same prompt of T2I-R1 to instruct the base model to generate a zero-shot semantic-level CoT, which we visualize in the figure and provide a comparison of the semantic-level CoT generated by T2I-R1.

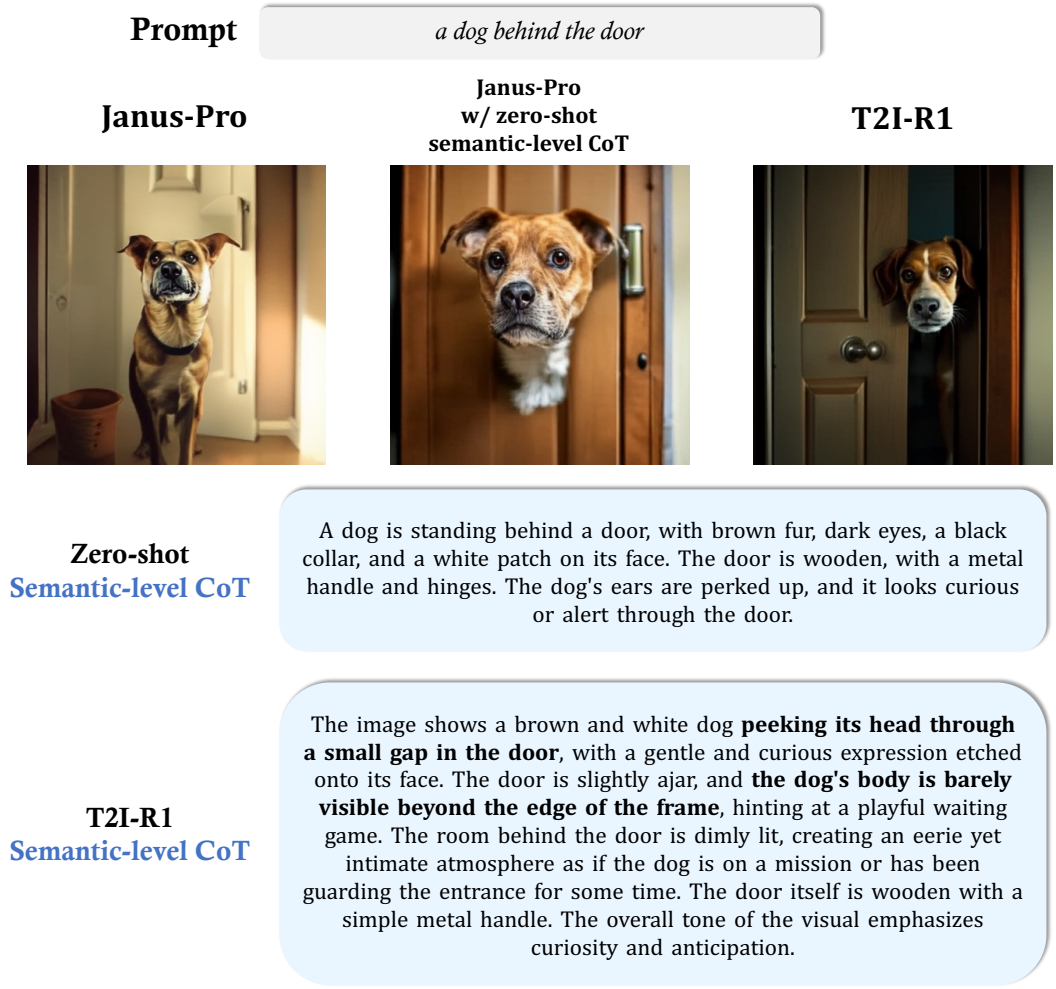

Figure 11: **Visualization Results of Semantic-level CoT.** We provide the image generation results of the same prompt from three settings: base model, base model with zero-shot semantic-level CoT, and T2I-R1. For the setting of base model with zero-shot semantic-level CoT, we use the same generation pipeline of T2I-R1 directly on the base model. We employ the same prompt of T2I-R1 to instruct the base model to generate a zero-shot semantic-level CoT, which we visualize in the figure and provide a comparison of the semantic-level CoT generated by T2I-R1.

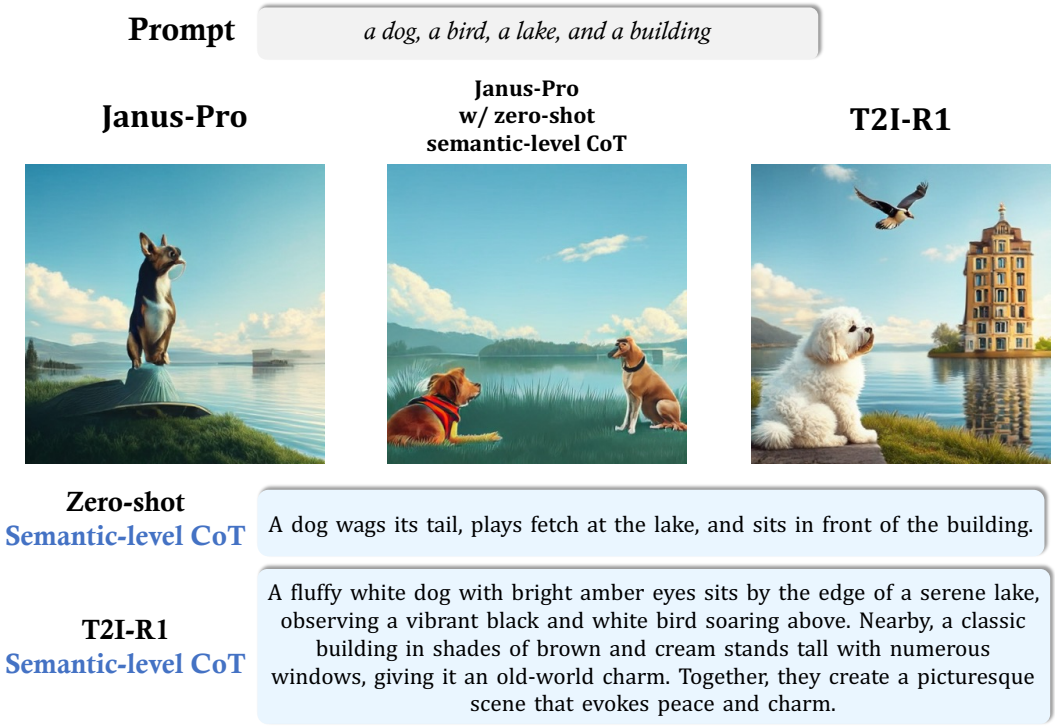

Figure 12: **Visualization Results of Semantic-level CoT.** We provide the image generation results of the same prompt from three settings: base model, base model with zero-shot semantic-level CoT, and T2I-R1. For the setting of base model with zero-shot semantic-level CoT, we use the same generation pipeline of T2I-R1 directly on the base model. We employ the same prompt of T2I-R1 to instruct the base model to generate a zero-shot semantic-level CoT, which we visualize in the figure and provide a comparison of the semantic-level CoT generated by T2I-R1.

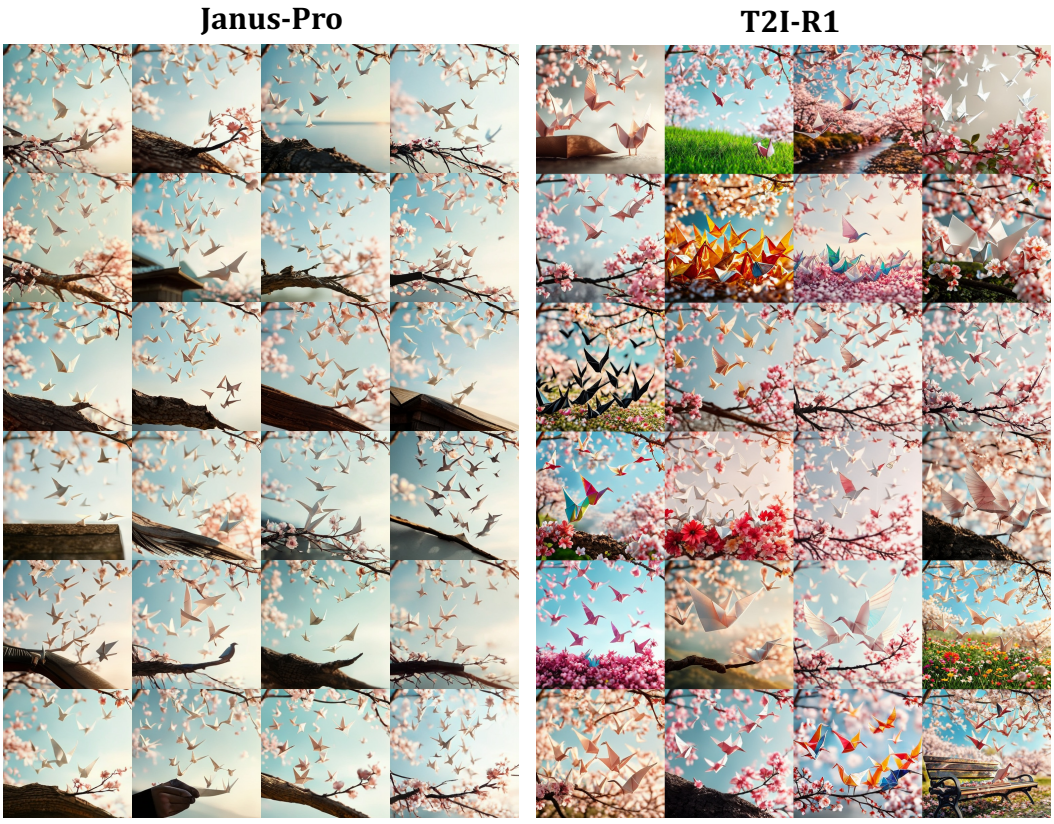

Figure 13: **More Visualization Result of the Image Diversity of a Single Prompt.** We showcase the result of the baseline model, Janus-Pro, and T2I-R1.

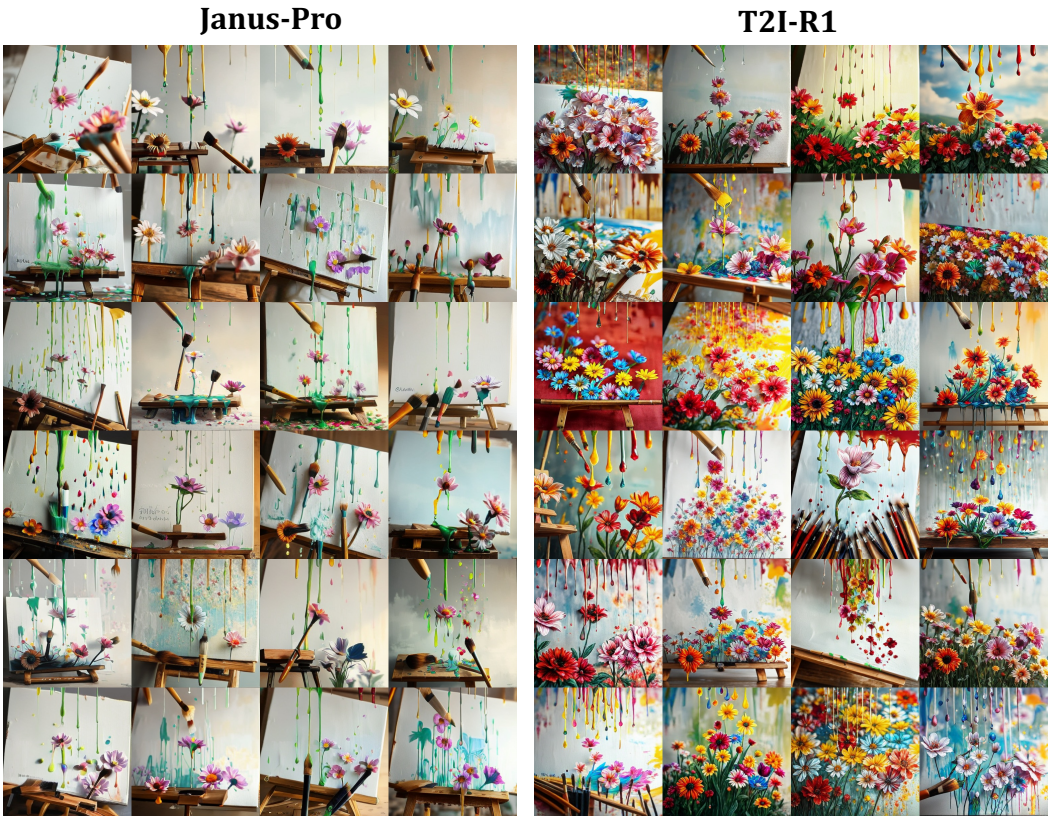

Figure 14: **More Visualization Result of the Image Diversity of a Single Prompt.** We showcase the result of the baseline model, Janus-Pro, and T2I-R1.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect this paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations of the work in Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see Section Experiments, Appendix, and the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Section Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We conduct experiments only once and report the accuracy of the best model, and it would be too computationally expensive to conduct the pre-training multiple times.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Section Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This work focuses on academic, publicly-available datasets including T2I-CompBench, WISE, and GenAI-Bench . This work is not related to any private or personal data, and there's no explicit negative social impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see Section Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Please see supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Please see Section Appendix.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.