Thinking-while-Generating: Interleaving Textual Reasoning throughout Visual Generation

Ziyu Guo[*1], Renrui Zhang[†*2], Hongyu Li[*3], Manyuan Zhang[†3], Xinyan Chen[2]
Sifan Wang, Yan Feng[3], Peng Pei[3], Pheng-Ann Heng[1]

CUHK [1]IMIXR & [2]MMLab [3]Meituan

Project Page: https://think-while-gen.github.io

Abstract

Recent advances in visual generation have increasingly explored the integration of reasoning capabilities. They incorporate textual reasoning, i.e., think, either before (as pre-planning) or after (as post-refinement) the generation process, yet they lack on-the-fly multimodal interaction during the generation itself. In this preliminary study, we introduce ***Thinking-while-Generating*** *(*TWIG*),* ***the first*** *interleaved framework that enables co-evolving textual reasoning throughout the visual generation process. As visual content is progressively generating, textual reasoning is interleaved to both guide upcoming local regions and reflect on previously synthesized ones. This dynamic interplay produces more context-aware and semantically rich visual outputs. To unveil the potential of this framework, we investigate three candidate strategies, zero-shot prompting, supervised fine-tuning (SFT) on our curated* TWIG*-50K dataset, and reinforcement learning (RL) via a customized* TWIG*-GRPO strategy, each offering unique insights into the dynamics of interleaved reasoning. We hope this work inspires further research into interleaving textual reasoning for enhanced visual generation. Code is released at: https://github.com/ZiyuGuo99/Thinking-while-Generating.*

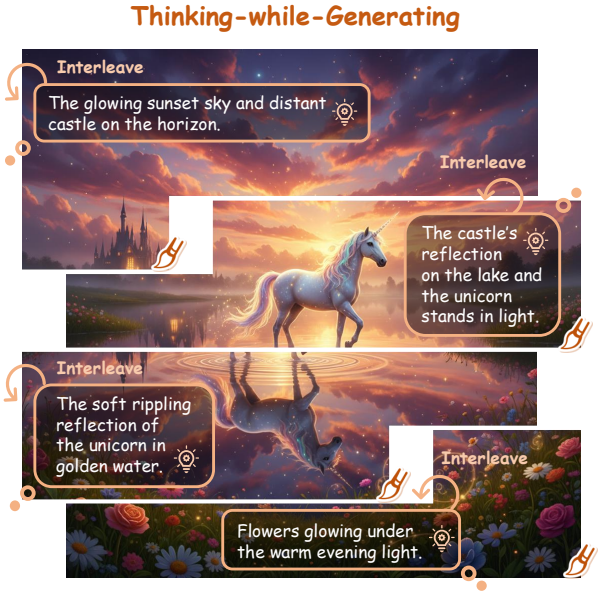

Figure 1. **Interleaving Textual Reasoning throughout Visual Generation.** Inspired by the image-interleaved reasoning in textual responses [8, 34, 45, 59], we reverse the modality flow and weave textual thoughts into the unfolding canvas, delivering on-the-fly guidance and reflection throughout synthesis.

1. Introduction

Visual generation have developed rapidly with diffusion [39, 41, 42] and autoregressive [7, 47, 54] models, enabling high-fidelity synthesis across diverse domains [33, 36, 57]. Despite impressive visual quality, today's generators often struggle with long-horizon composition, multi-entity relations, and adherence to nuanced textual instructions. Starting from '*Generation with CoT*' [17, 24, 48], a growing line of work explores *reasoning* as a remedy, typically injecting chain-of-thoughts in the language modality to assist visual synthesis.

Existing CoT-based approaches can be grouped by *where* the textual reasoning is applied, as compared in Figure 2:

- *Think before generation as a pre-planning aid.* Methods [11, 23, 28] first produce a structured or free-form plan, e.g., detailed captions, scene layouts, or object attributes and relations, and then condition the image generator on this plan. This improves global coherence and entity placement, but the plan is fixed once generation begins, limiting nuanced guidance and mid-course correction.
- *Think after generation as a post-refinement stage.* Meth-

[*]Equal Contribution [†]Project Lead

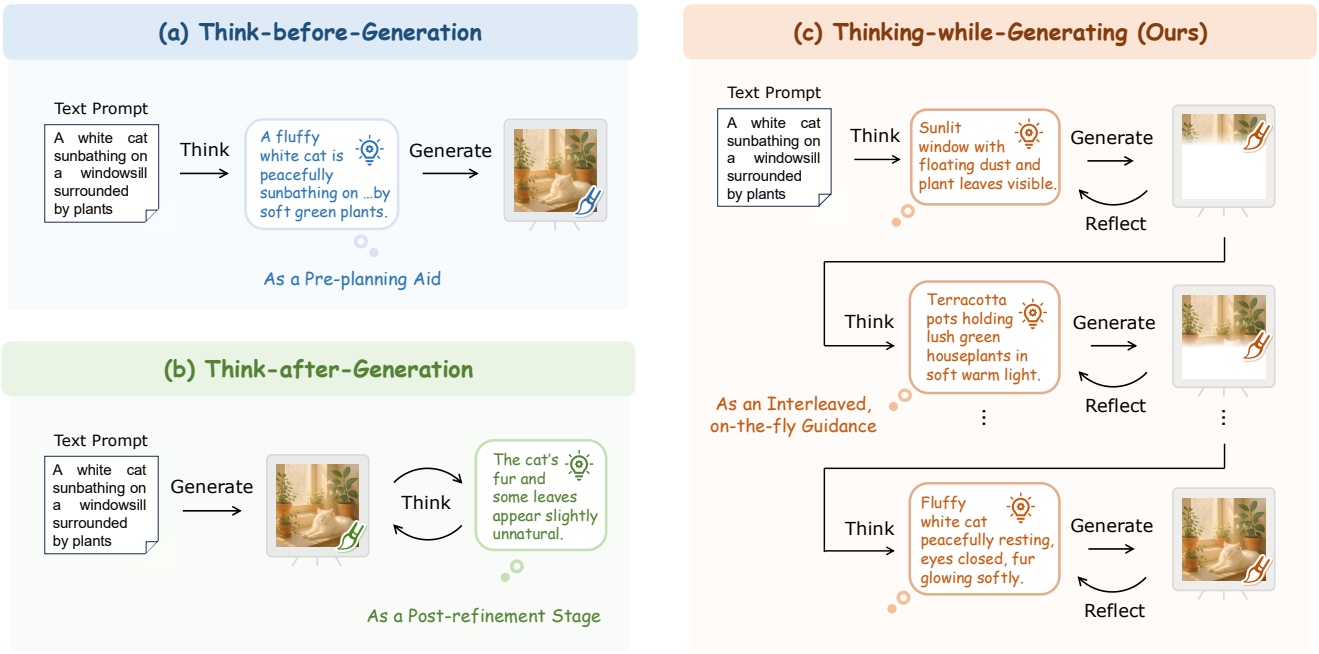

Figure 2. **Comparison of *Where* the Textual Reasoning is Applied in Visual Generation:** (a) *Think-before-Generation* [11, 23, 28] injects a pre-planning thought prior to the synthesis, limiting fine-grained control and later correction; (b) *Think-after-Generation* [17, 27, 38] verifies and revise the image once it is complete, lacking nuanced, timely adjustment with extra inference cost; (c) Our *Thinking-while-Generating* interleaves thoughts and reflections throughout the synthesis, providing on-the-fly. co-evolving guidance.

ods [17, 27, 61] synthesize the entire image first, and then elicit textual feedback via self-critique or external verifiers, iteratively revising the visual errors. These approaches help with local fixes and attribute binding, but reasoning is only loosely coupled to the synthesis trajectory without fine-grained, timely revision and, importantly, incur additional, costly extra inference rounds.

Given these limitations in visual generation, we note a complementary trend in visual understanding: recent large multimodal models (LMMs) [9, 13, 26, 58, 59] perform image-text interleaved reasoning, adaptively weaving intermediate visual evidence (e.g., detected objects, zoomed-in regions, or tagged images) into textual CoTs to improve interpretation and analysis. Inspired by this paradigm, we pose a natural question: as illustrated in Figure 1, *Can we invert the flow and interleave text into the intermediate visual generation process, providing on-the-fly, co-evolving reasoning that guides synthesis as it unfolds?*

In this preliminary study, we present *the first* interleaved framework for visual generation that keeps textual reasoning in the loop, termed as ***Thinking-while-Generating*** (TwiG), as compared in Figure 2 (c). As our approach is compatible with multiple models and task settings, for clarity and future extensibility, we adopt the unified understanding-generation LMM (ULM) [7, 52, 54, 60] with autoregressive generation paradigms, e.g., Janus-Pro [7], and experiment on text-to-image scenarios in our study.

Given a text prompt, the model first interprets the instruction and plans an optimal interleave schedule, i.e., how many steps to use and how to partition the canvas into local regions for progressive synthesis. While generating each region, the model conducts on-the-fly textual reasoning and grounds its thoughts in the current partial visual state. This interleaved *think* step serves two roles: *(i)* it produces nuanced guidance for the upcoming synthesis, and *(ii)* it critiques and reflects on the previously generated content. In this way, textual reasoning co-evolves with the visual modality, providing detailed, step-by-step directives. The image can be dynamically revised and precisely steered as it unfolds within a single generative trajectory.

We consider three candidate routes for *Thinking-while-Generating*, and investigate which, if any, proves effective:

- *Can zero-shot prompting alone achieve the goal?* We craft interleave-aware prompts to directly elicit global plans and reasoning thoughts. This route reveals the latent capacity of ULMs to self-organize interleaved reasoning without parameter updates, but can suffer instability.
- *Does supervised fine-tuning (SFT) benefit the performance?* We categorize the understanding and generation process into nine subtasks, and curate a dataset, TwiG-50K, for fine-tuning ULM, aiming to improve instruction adherence and reduce visual hallucination.
- *Will reinforcement learning (RL) further unlock its potential?* We optimize the interleaved reasoning policy of ULM via a customized GRPO [44] algorithm, TwiG-

GRPO, to push the performance boundary, investigating different RL approaches and reward designs.

Our experiments indicate that the ULM itself exhibits strong zero-shot capability for *Thinking-while-Generating*. With carefully designed prompts, it substantially improves Janus-Pro on T2I-CompBench(++) [19, 20] without additional training. Building on this, SFT with TWIG-50K provides further modest yet consistent gains, leading to more stable behavior compared with the zero-shot baseline. Finally, optimization with our TWIG-GRPO algorithm yields considerable improvements, underscoring the value of RL for deciding *when* to think, *what* to say, and *how* to refine. Taken together, these findings, though preliminary, are informative: they demonstrate the feasibility of interleaving textual reasoning during generation, and highlight this direction as a promising avenue for advancing visual synthesis.

It is worth noting that two relevant concurrent works, IRG [21] and Uni-CoT [38], attempt to 'interleave' reasoning with generation, but still treat the visual systhesis process as a monolithic block, like a combination of *think-before-generation* and *think-after-generation*. They are well-performed with unique insights, but not truly interleaving reasoning within the generative process itself, limiting the granularity and controllability.

2. *Thinking-while-Generating*

In Section 2.1, we first introduce the design scope and applicability of *Thinking-while-Generating*. Then, in Section 2.2, we present its overall pipeline and core components of the framework in detail.

2.1. Scope and Applicability

Aiming for generalization and extensibility, *Thinking-while-Generating* is conceptually compatible with diverse settings along the following three axes:

- **System Architecture.** The framework can be instantiated either as (i) a pipeline that couples a text-to-image model [2, 10, 39] with an LMM [1, 30, 56], where the LMM specializes in producing interleaved reasoning for the text-to-image outputs; or (ii) a ULM [7, 54, 60] that performs textual reasoning and visual generation within a single backbone.
- **Generation Paradigm.** The framework is applicable for visual generation with diffusion [12, 31, 40], discrete diffusion [4, 54, 55], and autoregressive models [7, 46, 47]. For continuous diffusion models, textual thoughts are interleaved at selected denoising steps; for discrete diffusion and autoregressive models, thoughts are inserted between segments of visual tokens to guide upcoming spans.
- **Task Scenarios.** The framework applies beyond T2I, e.g., image-to-image [3, 18, 57], text-to-video [14, 16, 37, 49], text-to-3D [15, 29, 36], and related generative tasks: as long as an LMM (or ULM) can provide reasoning thoughts for the target modality, they can be interleaved to steer generation.

As a preliminary study, we adopt a *single* ULM with an autoregressive generation paradigm (e.g., Janus-Pro [7]) for clarity of exposition, promising headroom, and end-to-end training efficiency. We denote its understanding forward pass by ULM_u and the generation forward pass by ULM_g.

2.2. Framework Overview

Figure 3 presents the overall *Thinking-while-Generating* (TWIG) framework, which interleaves textual reasoning with visual generation through three schemes: *when* to think, *what* to say, and *how* to refine.

***When* to Think (Scheduling).** Given an input prompt T, ULM_u first determines an interleaved reasoning schedule, denoted as $\mathcal{S} = \{\mathcal{V}_k\}_{k=1}^{K}$, according to:

$$\mathcal{S} = \mathrm{ULM}_u(T),$$

where each $\mathcal{V}_k$ denotes a target visual region at which reasoning is applied (e.g., token spans in autoregressive and discrete diffusion models, or timestep windows in continuous diffusion models). This decouples the generation process into smaller, more controllable sub-tasks guided by the interleaved textual reasoning. Scheduling can be static (fixed K, uniform spacing) or adaptive (variable K, content-dependent $\mathcal{V}_k$). In Section 3.1, we investigate different schedules and find that a static schedule with $K = 3$ performs the best, based on the heuristic that most images consist of three semantic components: upper background, central content, and lower background. Additionally, current capabilities of ULM_u are limited in reliably generating well-structured adaptive schedules, which remains a future work.

***What* to Say (Reasoning Content).** At each scheduled reasoning point, ULM_u provides a textual thought τ_k intended to guide the generation of the visual region $\mathcal{V}_k$. This thought serves as a localized sub-prompt exclusively targeted at $\mathcal{V}_k$, offering finer-grained guidance and alignment than prior *think-before-generation* approaches. The generation of τ_k is conditioned on three elements, i.e., the input prompt T, the previous thoughts $\{\tau_j\}_{j<k}$, and the visual content generated for prior regions $\{\mathcal{V}_j\}_{j<k}$, formulated as:

$$\tau_k = \mathrm{ULM}_u(T,\ \{\tau_j\}_{j<k},\ \{\mathcal{V}_j\}_{j<k}).$$

This allows τ_k to incorporate accumulated contextual information and to plan appropriately for the next visual segment. Subsequently, ULM_g synthesizes the target region $\mathcal{V}_k$, conditioned on all reasoning thoughts and the visual content

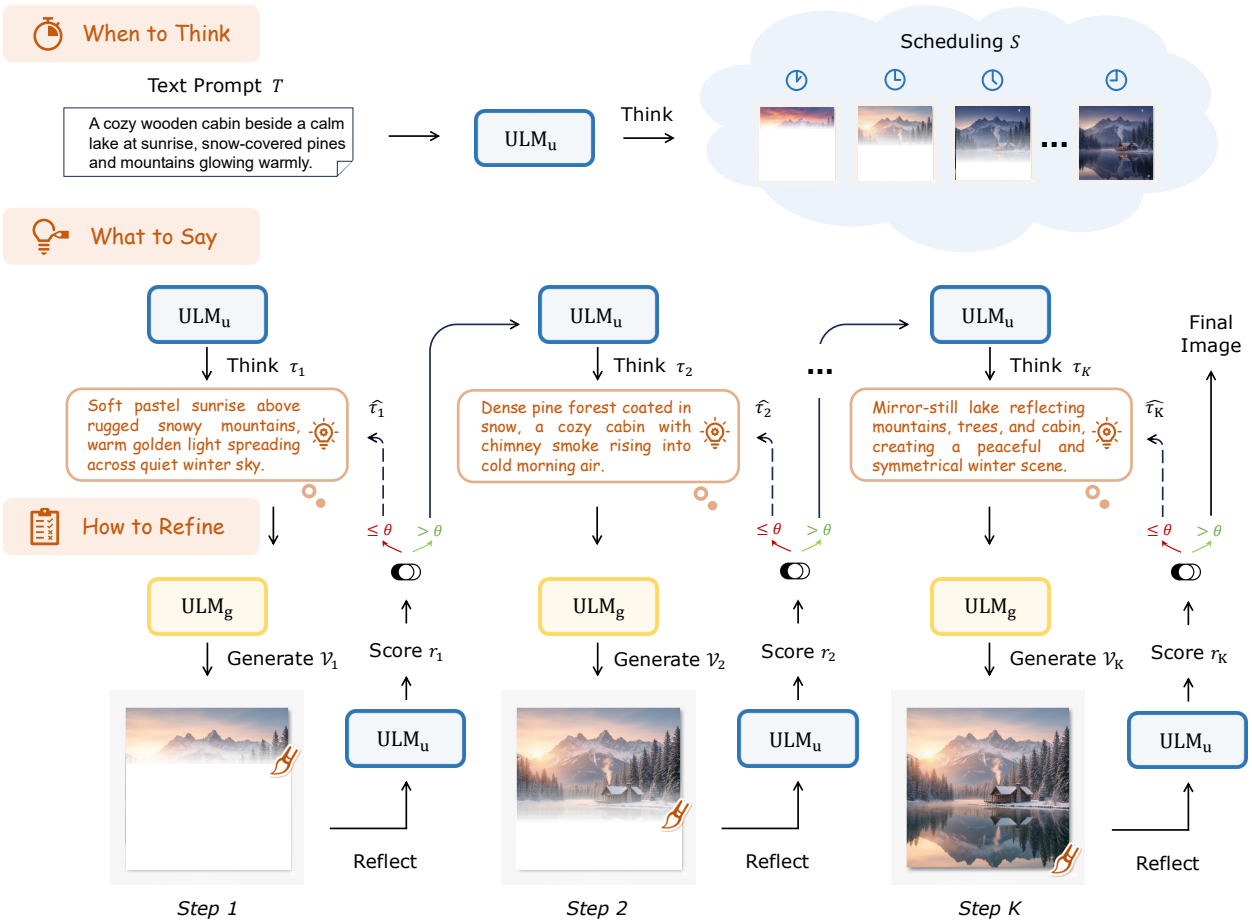

Figure 3. **Overall Pipeline of *Thinking-while-Generating*.** The framework comprises three components: *When to Think* for globally determining the interleaved generation schedule; *What to Say* for producing the step-by-step textual thought as fine-grained guidance; and *How to Refine* for a region-level reflection on the current canvas with optional corrective updates. ULM_u and ULM_g denote to apply a single ULM for understanding and generation, respectively.

produced up to by:

$$\mathcal{V}_k = \text{ULM}_g(\{\tau_j\}_{j\leq k},\ \{\mathcal{V}_j\}_{j<k}).$$

It is important to note that ULM_g ***is only required to possess text-to-image capabilities, no need for image-to-image functionality***. This is because the visual context $\{\mathcal{V}_j\}_{j<k}$ is not provided as image input to the model. Instead, we directly extend the textual pre-context from $\{\tau_j\}_{j<k}$ to $\{\tau_j\}_{j\leq k}$ at the beginning of the token sequence, while preserving the generated visual content $\{\mathcal{V}_j\}_{j<k}$ unchanged at the end of the sequence. This modification preserves the autoregressive generation process within a single trajectory, without introducing discontinuities or new generation rounds, as illustrated in Figure 4 (a).

***How* to Refine (Reflection).** After generating each visual region $\mathcal{V}_k$, we allow ULM_u to perform an immediate, region-level revision step that couples visual critique and an optional correction process. This enables finer-grained corrections while significantly reducing computational cost compared to prior *think-after-generation* approaches that conduct global post-revision. Before producing the next reasoning thought τ_{k+1}, ULM_u first generates a reflection tuple $c_k = (r_k, \hat{\tau}_k)$, given all the generated textual and visual contents as:

$$c_k = \text{ULM}_u(T,\ \{\tau_j\}_{j\leq k},\ \{\mathcal{V}_j\}_{j\leq k}),$$

where $r_k \in [0, 100]$ is an integer representing the critic score assigned to the current region $\mathcal{V}_k$, and $\hat{\tau}_k$ is a revised sub-caption intended for potential correction. The score r_k evaluates the semantic alignment and visual coherence of $\mathcal{V}_k$ with respect to its guiding prompt τ_k. If r_k exceeds a predefined threshold θ, the model proceeds directly to generate the next reasoning thought without revision. Otherwise, a local reflection is triggered to refine only the current sub-region, guided by $\hat{\tau}_k$, as defined by:

$$\hat{\mathcal{V}}_k = \text{ULM}_g(\{\tau_j\}_{j<k},\ \hat{\tau}_k,\ \{\mathcal{V}_j\}_{j<k}).$$

This localized corrective mechanism mitigates the accumulation of visual misalignments with timely revision. Likewise,

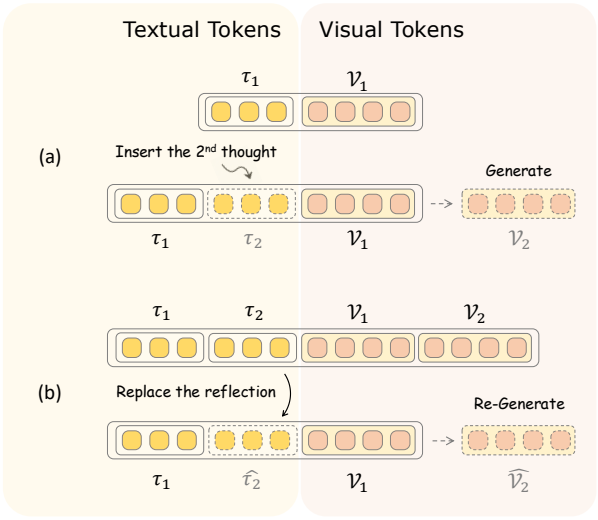

Figure 4. **Illustration of Interleaved Token Sequence:** (a) In *What to Say*, the textual pre-context extends from $\{\tau_j\}_{j<k}$ to $\{\tau_j\}_{j\leq k}$ ($K = 2$), guiding the generation of the next $\mathcal{V}_k$ while leaving the earlier $\{\mathcal{V}_j\}_{j<k}$ untouched; (b) In *How to Refine*, the thought τ_k is revised to $\hat{\tau}_k$, and only the local region $\hat{\mathcal{V}}_k$ is re-generated to replace $\mathcal{V}_k$. Neither operation requires the ULM to possess image-to-image capabilities, and both preserve a single text-to-image generation trajectory without launching a fresh pass or full re-generation.

as presented in Figure 4 (b), we directly update the textual pre-context from τ_k to the revised $\hat{\tau}_k$, and re-generate only the local part $\hat{\mathcal{V}}_k$ to replace $\mathcal{V}_k$ at the end of the token sequence, which also preserves a single trajectory without requiring the costly full re-generation.

In sum, *Thinking-while-Generating* (i) first schedules a number K of interleaved reasoning points (*when*); then for each $k = 1, \ldots, K$, (ii) produces a textual thought that locally steers the next visual update (*what*); and (iii) performs a region-level reflection with optional correction (*how*). The loop of (ii) and (iii) preserves a single generative trajectory, enabling on-the-fly guidance and precise local revision.

3. Implementation Exploration

In this section, we implement three candidate approaches for *Thinking-while-Generating*: zero-shot prompting (3.1), supervised fine-tuning (3.2), and reinforcement learning (3.3). We present experimental results that highlight their respective strengths. Please refer to detailed experimental settings and visualizations in the Supplementary Material.

3.1. Zero-shot Prompting

Prompt Customization. To elicit satisfactory zero-shot *Thinking-while-Generating*, we meticulously design a series of interleave-aware prompts for ULM, corresponding to the three components described in Section 2.2. Please refer to the final prompt templates in the Supplementary Material.

Table 1. **Zero-shot Experiments** of *Thinking-while-Generating* on T2I-CompBench [19]. We denote our zero-shot model as TwiG-ZS, and mark the improvement over the baseline, Janus-Pro-7B [7]. Panels *(a)*, *(b)*, *(c)*, and *(d)* present four ablation studies.

Setting	Attribute Binding			Object Relationship		Complex↑
	Color↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	
v.s. Baseline						
Janus-Pro-7B [7]	63.59	35.28	49.36	20.61	30.85	35.59
TwiG-ZS	73.11	41.55	64.77	21.98	30.90	36.65
Improve	*+9.52*	*+6.27*	*+15.41*	*+1.37*	*+0.05*	*+1.06*
(a) Where the Textual Reasoning is Applied						
Think-before-Gen.	65.12	36.20	51.05	20.88	30.82	35.92
Think-after-Gen.	64.72	37.95	50.62	21.05	30.87	36.27
Thinking-while-Gen.	73.11	41.55	64.77	21.98	30.90	36.65
(b) Interleaved Reasoning Step						
$K = 2$	72.79	42.26	64.64	21.97	30.89	36.83
$K = 3$	73.11	41.55	64.77	21.98	30.90	36.65
$K = 4$	72.95	41.90	64.70	22.03	31.10	37.02
(c) How to Partition $\mathcal{V}_k$ in Space						
Uniform Spacing	73.11	41.55	64.77	21.98	30.90	36.65
Adaptive Spacing	72.43	40.88	63.92	21.67	30.88	36.33
(d) Whether to Perform Reflection						
w/o Reflection	73.11	41.55	64.77	21.98	30.90	36.65
1-round Reflection	73.90	46.02	66.10	24.50	30.81	37.04
2-round Reflection	73.68	45.72	66.02	24.42	30.88	36.92

- For *when* to think, we prompt the model to adopt a global view, sketching the image's high-level semantics and structure step by step from the input prompt. For an adaptive schedule, we additionally prompt the model to output the relative ratios of visual parts across the canvas.
- For *what* to say, we guide the model to focus strictly on the local region currently being generated while maintaining coherence with previously generated visual and textual context. We discourage any spatial-anchor tokens; the model should produce only the descriptive content.
- For *how* to refine, we prompt the model to provide a critic score evaluating along five criteria (color accuracy, object completeness, detail richness, spatial relationships, and visual coherence), ensuring a consistent standard across cases. The template enforces that any revision is local and does not contradict validated prior regions.

Experiments and Analysis. In Table 1 (top), we present the performance of our zero-shot model, TwiG-ZS. We observe that our carefully designed prompts yield surprisingly strong improvements over the baseline, significantly surpassing Janus-Pro-7B [7] across multiple dimensions. This highlights the potential of our framework and its natural applicability within current ULMs, making the zero-shot variant a strong foundation for subsequent SFT and RL. By default, we adopt an interleaved schedule with $K = 3$ and uniform spacing, and permit at most one round of reflection. We conduct four ablations:

- **Ablation *(a)*:** *Thinking-while-Generating* versus *Think-before/after-Generation* under identical zero-shot settings. Interleaving provides nuanced, on-the-fly guidance rather

Table 2. **SFT Experiments** of *Thinking-while-Generating* on T2I-CompBench [19]. We denote our fine-tuned model as TwiG-SFT, and mark the improvement over TwiG-ZS. Panel *(a)* ablates the varying proportions of thinking (*T*), generation (*G*), and reflection (*R*) data in TwiG-50K. Panel *(b)* reports the standard deviation (*Std*) across random seeds to assess stability.

Model / Setting	Data *T* / *G* / *R*	Attribute Binding Color↑	Shape↑	Texture↑	Object Relationship Spatial↑	Non-Spatial↑	Complex↑
v.s. Baseline							
Janus-Pro-7B [7]	–	63.59	35.28	49.36	20.61	30.85	35.59
TwiG-ZS	–	73.11	41.55	64.77	21.98	30.90	36.65
TwiG-SFT		74.58	52.42	67.95	27.02	31.24	38.22
Improve	–	*+1.47*	*+10.87*	*+3.18*	*+5.04*	*+0.34*	*+1.57*
(a) Effect of Training Data Composition							
Think-heavy		73.38	50.92	66.47	26.08	30.97	37.68
Gen-heavy		74.12	51.77	67.28	26.58	31.09	37.94
Think-Gen-equal		74.58	52.42	67.95	27.02	31.24	38.22
Reflect-lite		72.76	49.75	65.93	26.36	30.92	37.21
Reflect-heavy		71.88	48.98	65.05	25.62	30.84	36.87
(b) Stability across 5 Random Seeds							
TwiG-ZS *Std*↓	–	0.82	0.70	0.76	0.45	0.38	0.91
TwiG-SFT *Std*↓		0.65	0.59	0.61	0.40	0.36	0.80

Table 3. **RL Experiments** of *Thinking-while-Generating* on T2I-CompBench [19]. We denote our reinforced model with GRPO [43] as TwiG-RL, and mark the improvement over the TwiG-SFT. Panels *(a)* and *(b)* present the results of two ablation studies.

Setting	Attribute Binding Color↑	Shape↑	Texture↑	Object Relationship Spatial↑	Non-Spatial↑	Complex↑
v.s. Baseline						
Janus-Pro-7B [7]	63.59	35.28	49.36	20.61	30.85	35.59
TwiG-ZS	73.11	41.55	64.77	21.98	30.90	36.65
TwiG-SFT	74.58	52.42	67.95	27.02	31.24	38.22
TwiG-RL	82.49	61.28	73.19	34.06	31.99	40.31
Improve	*+7.91*	*+8.86*	*+5.24*	*+7.04*	*+0.75*	*+2.09*
(a) TwiG-*GRPO Strategy*						
ULM_g-GRPO	80.12	59.87	72.01	32.47	31.30	39.88
ULM_u-GRPO	78.36	57.94	70.68	30.93	31.27	39.63
TwiG-GRPO	82.49	61.28	73.19	34.06	31.99	40.31
(b) Reward Model Ensemble						
Human Preference	79.83	60.97	71.35	20.68	30.53	38.71
+ Object Grounding	80.44	60.01	73.79	25.84	31.15	39.94
++ VQA Consistency	80.87	59.29	74.26	30.05	31.41	39.52
+++ LMM Alignment	82.49	61.28	73.19	34.06	31.99	40.31

than only pre-planning or post-refinement, and consistently outperforms the alternatives.

- **Ablation *(b):*** Number of interleaved reasoning steps under a uniform schedule. We find $K = 3$ is optimal, aligning with the heuristic that many images decompose into three semantic components: upper background, central content, and lower background.
- **Ablation *(c):*** Adaptive scheduling of interleaved spacing. Despite exploring multiple prompting strategies, current ULMs struggle to reliably follow such instructions, leading to unstable or poorly structured adaptive schedules.
- **Ablation *(d):*** Effectiveness of reflection during reasoning. A single reflection round corrects misalignments and improves performance across aspects; however, conducting two rounds brings no further gains, likely limited by the critique–and–revision capacity of zero-shot ULMs.

3.2. Supervised Fine-tuning

SFT Task Formulation. Building on the zero-shot baseline, we investigate whether SFT can enhance the capabilities. We decompose the *Thinking-while-Generating* process into nine supervised tasks that mirror the inference loop, using a fixed number of three reasoning steps. These comprise three thinking targets for ULM_u (upper/central/lower thoughts), three reflection targets for ULM_u (three scores with revised thoughts), and three generation targets for ULM_g (three visual regions). This enables the model to learn structured reasoning, localized reflection, and region-wise generation in an interleaved, context-aware manner.

TwiG-50K Dataset. To support the task formulation, we curate a high-quality dataset termed TwiG-50K. The construction process comprises multiple stages of synthetic supervision using advanced commercial models.

- For *what* to say (∼17K, three tasks), we source 5.5K text prompts from the training split of T2I-CompBench [19], and adopt GPT-4o [22] to generate stepwise sub-captions that segment the image into three coherent parts (upper background, central content, lower background). These sub-captions are concatenated and fed to GPT-4o-Image [22] to synthesize images that are semantically consistent with the specified divisions. We then filter low-quality instances and organize them into interleaved formats aligned with the *Thinking-while-Generating* protocol. Note that, since the reasoning step count is fixed to three, we do not collect supervision data for *when* to think.
- For *how* to refine (∼17K, three tasks), building on the interleaved samples above, we construct three visual understanding tasks focused on critique and revision. GPT-4o is prompted to evaluate each region by assigning a critic score along five criteria (the same as zero-shot settings) and to provide a revised sub-caption that addresses deficiencies identified by the critique. If the original image attains a high score, the revised thought simply repeats, a case that may not trigger re-generation during inference.
- To enhance the generation capability of ULM_g (∼16K, three tasks), we construct interleaved visual generation data from the image–sub-caption pairs obtained in the *when/what* stage. Each training instance conditions the generation of region $\mathcal{V}_k$ on cumulative reasoning thoughts $\{\tau_j\}_{j\le k}$ and previously generated visual contents $\{\mathcal{V}_j\}_{j<k}$. Note that this remains text-to-image supervision to preserve a single generation trajectory (not image-to-image), augmented with a visual pre-context.

Experiments and Analysis. In Table 2 (top), we present the performance of our fine-tuned model, TwiG-SFT. Relative to the zero-shot baseline (TwiG-ZS), SFT delivers modest and reliable gains across benchmarks, with the largest

Table 4. **Performance Comparison on T2I-CompBench++ [20].** The best and the second-best scores are highlighted.

Model	Attribute Binding			Object Relationship			Numeracy↑	Complex↑
	Color↑	Shape↑	Texture↑	2D-Spatial↑	3D-Spatial↑	Non-Spatial↑		
Current Generative Models								
Show-o [54]	56	41	46	20	-	30	-	29
SD-XL-base-1.0 [35]	58.79	46.87	52.99	21.31	35.66	31.19	49.91	32.37
Attend-and-Excite [5]	64.00	45.17	59.63	14.55	32.22	31.09	47.73	34.01
PixArt-α [6]	66.90	49.27	64.77	20.64	-	31.97	-	34.33
GoT [11]	65.51	50.08	58.36	24.57		31.13	-	37.54
Show-o + PARM [17]	75	56	66	29	-	31	-	37
FLUX.1 [25]	74.07	57.18	69.22	28.63	38.66	31.27	61.85	37.03
Emu3 [51]	75.44	57.06	71.64	-	-	-	-	-
T2I-R1 [23]	81.30	58.52	72.43	33.78	-	30.90	60.97	39.93
Thinking-while-Generating								
Janus-Pro-7B [7] (Baseline)	63.59	35.28	49.36	20.61	32.94	30.85	41.32	35.59
TwiG-ZS	73.11	41.55	64.77	21.98	33.68	30.90	50.01	36.65
TwiG-SFT	74.58	52.42	67.95	27.02	35.57	31.24	51.70	38.22
TwiG-RL	82.49	61.28	73.19	34.06	38.87	31.99	61.93	40.31

improvements on *Shape* and *Spatial* categories. This demonstrates the effectiveness of our fine-tuning recipe and the curated TwiG-50K dataset. By default, we inherit the optimal model settings from TwiG-ZS, and adopt a balanced data mixture with equal thinking and generation tasks. We further provide two analyses:

- **Ablation *(a)*:** Effect of data composition from TwiG-50K. Balancing thinking (*T*) and generation (*G*) provides the best trade-off and strengthens *Thinking-while-Generating* from both sides. However, adding reflection data (*R*) degrades the results, where the thoughts become longer and over-corrections appear more frequently. This suggests that, TwiG-ZS already exposes most of the model's reflection proficiency, and oversupplying *R* diverts capacity away from learning stable *T* and *G* behaviors. Although the reflection subset cannot contribute here, we hope it will facilitate future research on critique-and-revise training.
- **Comparison *(b)*:** Inference stability across five random seeds. We report the standard deviation (*Std*) over different runs, and observe that SFT notably tightens dispersion compared to TwiG-ZS, indicating more predictable behavior. Qualitatively, SFT shortens verbose thoughts, curbs hallucinations, improves attribute persistence across adjacent regions, and reduces spurious reflection triggers near the decision threshold.

3.3. Reinforcement Learning

TwiG-GRPO Strategy. To further advance performance, we employ RL to enhance the interleaved reasoning. Specifically, we adopt the GRPO algorithm [43] with the training prompts from T2I-CompBench, and tailor it to our *Thinking-while-Generating* framework. Within this setup, the ULM performs multiple forward passes within a single rollout during GRPO training. A key design question is which components should be reinforced through the reward mechanism: all stages, or only the understanding or generation phases? We propose to reinforce all of them simultaneously through our TwiG-GRPO strategy. Concretely, we compute a single reward based on the final generated image and the input prompt, and utilize it as a shared reward to optimize the policies of every thinking, generation, and reflection pass jointly. This approach not only simplifies implementation (no need to compute rewards for each local visual subtask), but also enables consistent reinforcement across ULM_u and ULM_g, allowing global information to flow across different paths and thereby enhancing the overall synergy of the TwiG framework.

Reward Model Design. Since a high-quality image must satisfy multiple aspects (overall aesthetics, object attributes and relationships), we explore to combine complementary reward models for joint optimization and mitigating reward hacking [23]: (i) human preference scores (HPS v2 [53]), (ii) object grounding scores (GroundingDINO [32]), (iii) VQA consistency scores (GIT [50]), and (iv) LMM alignment scores (the fine-tuned ORM [17]). We utilize an unweighted average of the four reward model, and this simple strategy effectively leverages our framework's generality for RL gains.

Experiments and Analysis. In Table 3 (top), we present the performance of our reinforced model, TwiG-RL. Compared with TwiG-SFT, the initialization point, RL delivers substantial gains, e.g., exceeding +5%, across the three *Attribute Binding* categories and the *Spatial* category. This highlights the remaining headroom of the *Think-while-Generating* paradigm once a policy is guided in a right direction with an appropriate GRPO strategy and reward ensemble designs. In Table 4, we report the three TwiG

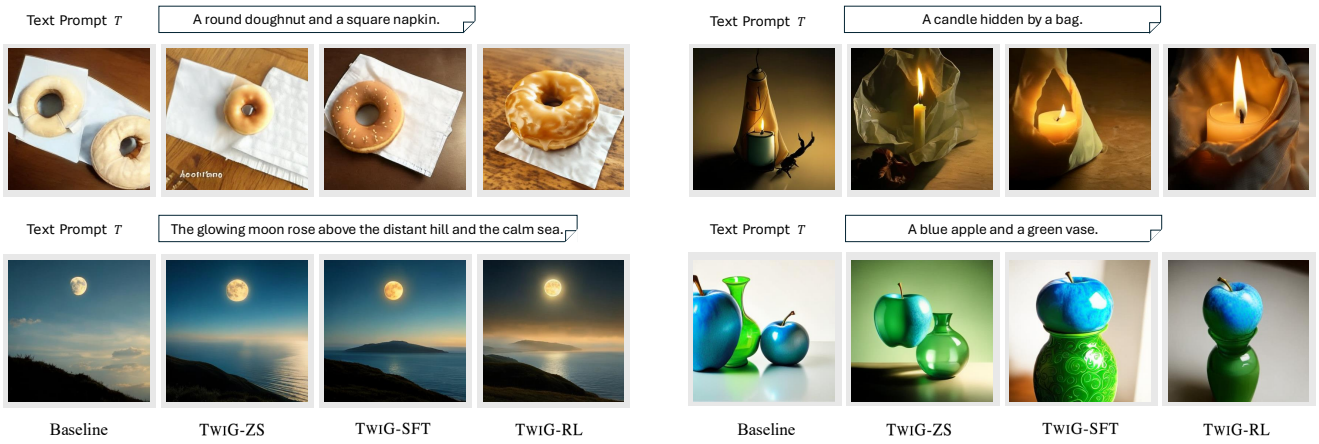

Figure 5. **Qualitative Comparison of TwiG Variants:** the baseline (Janus-Pro-7B [7]), TwiG-ZS, -SFT, and -RL. Our method demonstrates progressive improvements in compositional fidelity, object counting, and visual realism.

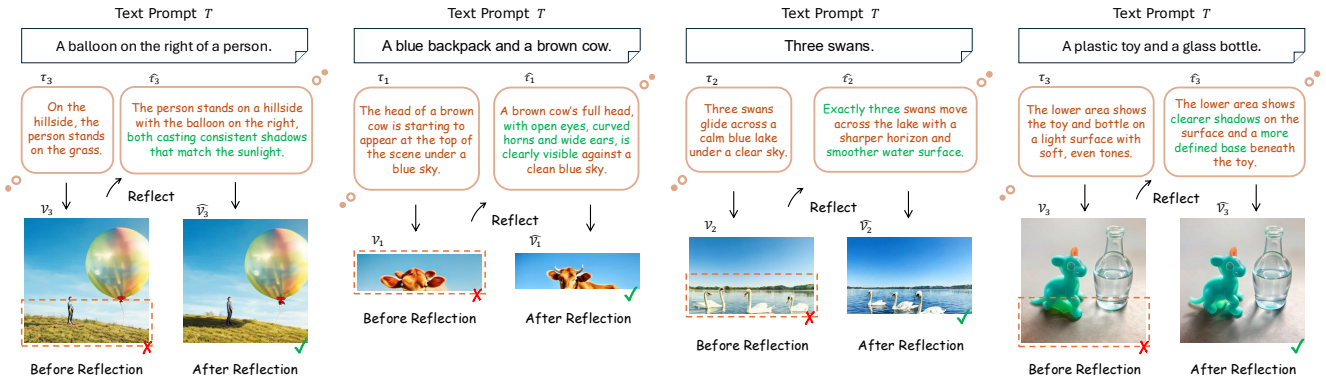

Figure 6. **The Reflection Capacity of TwiG-RL**. The reflection within our *Thinking-while-Generating* refines both semantic and visual consistency, e.g., improving spatial alignment, shadow coherence, and overall realism across diverse prompts.

approaches in comparison with current generative models on T2I-CompBench++ [20]. Our method offers a flexible trade-off between implementation efficiency (ZS) and competitive performance (RL), allowing practitioners to balance the cost and quality according to deployment needs. Furthermore, in Figures 5 and 6, we present three visualizations, i.e., illustrating the improvements across different variants, the reflection capability, and the image-text interleaved reasoning process, respectively, which highlight the qualitative effectiveness of our methods.

- **Ablation *(a)*:** Different strategies for GRPO algorithms. Our TwiG-GRPO jointly reinforces all (up to nine) local visual subtasks within a single rollout. We investigate to separately optimize the understanding-related tasks (thinking and reflection) and the generation-related tasks, each using the shared reward to update ULM_u and ULM_g, respectively. As compared, the separate enhancements fail to surpass the joint strategy, highlighting their complementary nature and mutual reinforcement. Only when combined under the full TwiG-GRPO strategy can the RL potential of the interleaved reasoning be fully realized.
- **Ablation *(b)*:** Ensemble of multiple reward models. We begin with a single HPS v2, and progressively incorporate other three rewards. HPS v2 primarily improves global aesthetics and stylistic coherence; GroundingDINO tightens entity presence and localization; GIT curbs instruction violations and strengthens attribute consistency; the fine-tuned ORM improves holistic text–image alignment. Adding components steadily improves performance, and the ensemble of four achieves the best overall balance.

4. Conclusion

In this paper, we introduce the *Thinking-while-Generating* (TwiG) paradigm, an interleaved framework that keeps textual reasoning in the loop during visual generation. Starting from carefully designed zero-shot prompts, then enhancing with SFT, and finally optimizing a policy via RL, our TwiG model learns to think, generate, and reflect within a single visual generation trajectory. We hope this paradigm may inspire future research to fully investigate the potential of interleaved visual generation schemes.

Acknowledgements

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project 14200824 and Project 14202125.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

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

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022. 3

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 7

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 7

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 1, 2, 3, 5, 6, 7, 8

[8] Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025. 1

[9] Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *arXiv preprint arXiv:2510.11718*, 2025. 2

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

[11] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025. 1, 2, 7

[12] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 3

[13] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025. 2

[14] Google DeepMind. Veo-3 technical report. Technical report, Google DeepMind, 2025. 3

[15] Ziyu Guo*, Renrui Zhang*, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3

[16] Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025. 3

[17] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 1, 2, 7

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[19] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 3, 5, 6

[20] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3, 7, 8

[21] Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025. 3

[22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[23] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 1, 2, 7

[24] Dongzhi Jiang, Renrui Zhang, Haodong Li, Zhuofan Zong, Ziyu Guo, Jun He, Claire Guo, Junyan Ye, Rongyao Fang, Weijia Li, et al. Draco: Draft as cot for text-to-image preview and rare concept generation. *arXiv preprint arXiv:2512.05112*, 2025. 1

[25] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 7

[26] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 2

[27] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. *arXiv preprint arXiv:2503.12271*, 2025. 2

[28] Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. *arXiv preprint arXiv:2503.19312*, 2025. 1, 2

[29] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 3

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[31] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer, 2022. 3

[32] Siyi Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023. 7

[33] OpenAI. Sora 2 system card. Technical report, OpenAI, 2025. 1

[34] OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, 2025. 1

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

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3

[37] Yu Qi, Xinyi Xu, Ziyu Guo, Siyuan Ma, Renrui Zhang, Xinyan Chen, Ruichuan An, Ruofan Xing, Jiayi Zhang, Haojie Huang, et al. Mme-cof-pro: Evaluating reasoning coherence in video generative models with text and visual hints. *arXiv preprint arXiv:2603.20194*, 2026. 3

[38] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025. 2, 3

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[43] Zhenming Shao, Jiayi Gu, Ziyang Wang, Liang Ding, Yi Wang, Yao Zhang, Shuming Tang, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces Group Relative Policy Optimization (GRPO) used for RL training. 6, 7

[44] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2

[45] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025. 1

[46] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

[47] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 3

[48] Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025. 1

[49] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[50] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 7

[51] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen

Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 7

[52] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 2

[53] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 7

[54] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 2, 3, 7

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

[56] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1, 3

[58] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 2

[59] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1, 2

[60] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 2, 3

[61] Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15329–15339, 2025. 2