DEBATE-DRIVEN GENERATION: MULTIMODAL AGENT DEBATE FOR ITERATIVE TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in text-to-image generation have placed growing emphasis on reasoning for better instruction following and refinement. Nevertheless, the need to transform sparse language into dense visual content makes contextual disambiguation and elaboration a central bottleneck beyond cross-modal alignment. Therefore, we introduce the Multimodal Agent Debate (MAD) framework that adapts multi-agent debate to iterative text-to-image refinement by addressing three gaps, namely evolving debate topics, the mismatch between semantic precision and execution feasibility, and the need for factual rather than reflective moderation. MAD maintains (i) an explicit roadmap of verifiable criteria grounded in the current image, (ii) uses a Regulator to decide whether debated instructions are executable by a downstream editor or require re-debate with pragmatic constraints, and (iii) employs a cross-modal Moderator to verify prompt satisfaction and produce verifiable success conditions. Experiments on GenEval, T2I-CompBench, and TIIF-Bench show consistent overall gains in instruction following across diverse image generators. In addition, MAD also improves instruction executability and reduces repetitive refinements, supporting MAD as a training-free refinement method for T2I. The source code will be released upon acceptance.

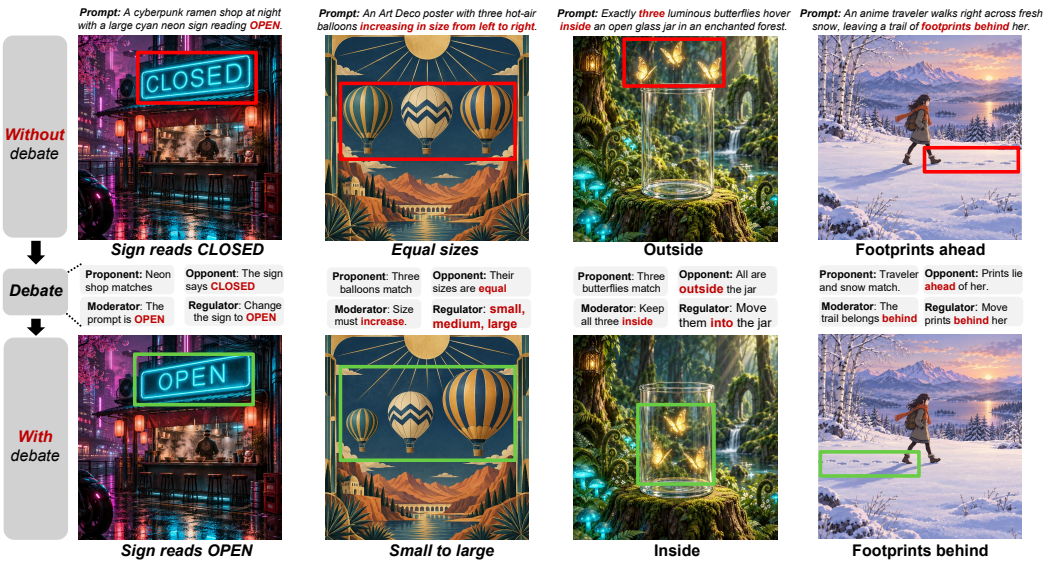

Figure 1: Qualitative examples of Multimodal Agent Debate (MAD). Through debate-guided refinement, MAD corrects errors across diverse scenarios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Image-to-Text (I2T) and Text-to-Image (T2I) generation constitute two pivotal directions in multimodal understanding. I2T has attained notable success, whereas T2I remains more challenging despite its promise. Both tasks can be framed as translation between vision and language, and the community has therefore concentrated on specialized mapping models, namely LLMs for I2T Zheng et al. (2024); Liang et al. (2025) and diffusion models for T2I Labs et al. (2025); Liu et al. (2025). A key distinction between them, however, lies in their information density. I2T maps information-dense images to sparser textual representations, whereas T2I performs the inverse mapping. From a linguistic perspective, translating sparse text into a dense visual format requires substantial contextual disambiguation and elaboration beyond mere cross-modal alignment. We argue that this inherent asymmetry is the primary and previously overlooked reason for the larger potential for improvement in T2I systems.

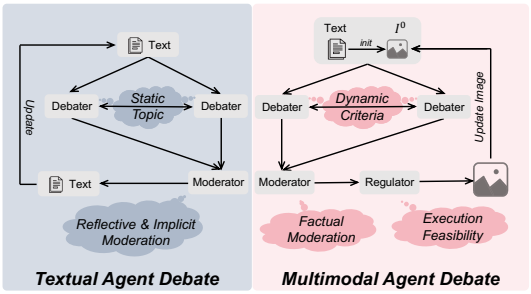

Figure 2: Comparison of textual agent debate and multimodal agent debate.

In real-world professional visual production, T2I Zhang et al. (2025a); Jiang et al. (2025) typically relies on a collaborative process between designers and art directors. They engage in discussion and iterative refinement to achieve the required contextual disambiguation and elaboration. This paradigm is closely aligned with multi-agent debate Hu et al. (2025); Du et al. (2023), in which multiple agents argue and collaborate to refine a shared output. Motivated by this alignment, we conduct a pilot study to examine whether a multimodal agent debate can provide robust contextual inference for T2I generation.

Applying a debate framework to Text-to-Image generation is nonetheless non-trivial. The primary challenge arises from three critical gaps between conventional textual debate and the proposed Multimodal Agent Debate (MAD) framework, which is a T2I-specific instantiation of debate shown in Fig. 2. We describe these gaps as follows.

- **Static vs. Dynamic Topics**: Unlike a textual debate with a fixed proposition, MAD addresses multiple evolving topics, such as color accuracy, object layout, and aesthetics, whose focus shifts at each generation step. To address this gap, a set of verifiable criteria is defined and structured as an iterative refinement roadmap. These criteria operationalize clarification by transforming a prompt into a sequence of progressively densified constraints. They specify which semantic aspects require explicit confirmation, what counts as visual evidence, and how an objection must be grounded in the generated image to be testable.

- **Semantic Fidelity vs. Execution Feasibility**: Debate agents excel at producing precise textual clarifications. However, such linguistic precision does not guarantee that the required edits are technically feasible for a vision model. To resolve this gap, a Regulator role is incorporated. The Regulator is positioned at the critical junction between language-based debate and visual execution. Its core function is to assess whether a debated instruction can be translated into an executable operation by the downstream Executor in light of the generation history. It then orchestrates the workflow and decides whether to proceed with execution or to initiate a re-debate under more pragmatic constraints.

- **Reflective vs. Factual Moderation**: In a textual debate, a judge agent monitors progress and terminates deliberation based on its reflection on implicit linguistic evidence. In MAD, the goal is instead factual and cross-modal, and it requires explicit verification that the image satisfies the prompt. To address this gap, a cross-modal Moderator is introduced. Its role is to make the termination decision factual by evaluating the current image against the prompt. It consolidates the entire debate into a concrete binary judgment of satisfaction together with a derived set of verifiable conditions. Consequently, MAD attains principled and evidence-based termination rather than relying on heuristics.

2 Related Work

2.1 Text-to-Image Generation

T2I Labs (2024); Podell et al. (2023); Labs et al. (2025); Zhang et al. (2024); Esser et al. (2024) generates images from textual input, and it provides a natural testbed for studying how sparse language signals are translated into dense visual content. Existing T2I models can be broadly organized into three paradigms, namely autoregressive generation, diffusion-family generation, and unified multimodal generation. Autoregressive models Wu et al. (2025) synthesize images by sequentially predicting discrete visual tokens. Diffusion-family models include both diffusion and flow-matching variants, and they generate images through iterative denoising or related continuous transport processes in latent space. Representative examples include the Stable Diffusion series, Hunyuan-DiT Li et al. (2024), SANA-1.5 Xie et al. (2025), and Kontext Labs et al. (2025). Unified multimodal generation models instead support both visual understanding and image generation within a shared architecture, such as the Janus series Ma et al. (2025) and Bagel Deng et al. (2025).

2.2 Reasoning for Image Generation and Editing

Recent work on reasoning for image generation and editing can be roughly grouped into two lines, namely CoT-based methods and multi-agent methods.

CoT-based methods introduce stepwise reasoning, verification, reflection, or reinforcement learning into the generation loop. ImageCoT and T2I-R1 strengthen generation through intermediate reasoning and iterative verification Guo et al. (2025); Xiao-Yong Wei (2013); Wei & Yang (2011); Jiang et al. (2025). GoT Fang et al. (2025) further extends reasoning-guided generation and editing to more complex settings. It improves semantic planning and prompt alignment, but its reasoning remains centered on a single model or controller. Multi-agent methods instead distribute reasoning across specialized roles, such as interpretation, planning, evaluation, and scene parsing. T2I-Copilot Chen et al. (2025a) employs multiple agents for prompt understanding, model selection, and quality evaluation, whereas MCCD Li et al. (2025) combines multi-agent scene parsing with compositional diffusion for complex generation. However, these systems are largely cooperative. They emphasize decomposition and coordination, and they do not formulate refinement as an adversarial multimodal debate over shared visual evidence.

Despite these differences, both lines share a common limitation in iterative refinement. Long reasoning chains can encourage overthinking and propagate early errors, which leads to accumulated drift rather than more accurate refinement.

2.3 Multi-agent Debate

Multi-agent debate frames reasoning as a role-based interaction among multiple LLM Chen et al. (2025b); Achiam et al. (2023); Grattafiori et al. (2024) and MLLM Bai et al. (2025); Wang et al. (2025); Cheng et al. (2025); Team et al. (2023) agents. Such debates can generally be divided into two types by modality, namely text-only debates and multimodal debates. Text-only debate Hu et al. (2025); Du et al. (2023); Sheng et al. (2025) is primarily proposed to mitigate the severe hallucinations that arise from LLM reasoning. Debate-Augmented RAG Hu et al. (2025) follows this design and introduces debate stages for retrieval and response generation, which reduces hallucinations by strengthening evidence selection and consistency. Multimodal debate applies the debate paradigm to visual understanding tasks Zheng et al. (2024); Liang et al. (2025). However, it remains underexplored for generative objectives such as text-to-image synthesis.

3 Method

The pronounced information-density mismatch between language and images induces confabulatory judgements and makes faithful instruction following difficult. We therefore treat T2I as an iterative cycle of initial generation, debate, revision under a bounded number of debate rounds, namely Multimodal Agent Debate (MAD). The framework of MAD is shown in Fig. 3.

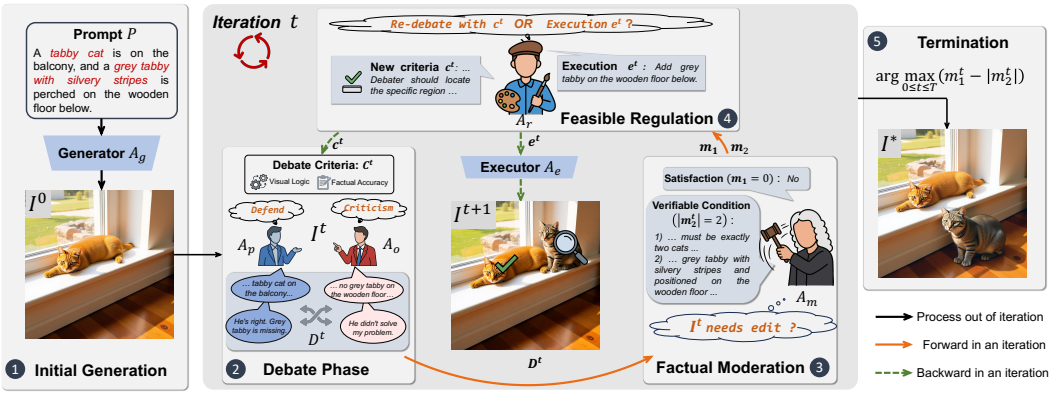

Figure 3: Overview of Multimodal Agent Debate. It illustrates five components: 1. initial image generation; 2. dynamic criteria debate phase; 3. factual moderation; 4. feasible regulation; 5. Iteration Termination.

3.1 Preliminary

We consider iterative text-to-image generation as the MAD system S. For each generation task instance, S is defined as a tuple of six components, namely a prompt P, an image state I, a set of N agents $\mathcal{A} = \{\mathcal{A}_i\}_1^N$, a family of multimodal debating functions $\mathcal{F}$, and a set of agent responses $\mathcal{R} = \{R_i\}_1^N$, and debate history records H^t.

Before debate begins, the generator A_g produces the initial image I^0 from the prompt P.

MAD then performs at most T outer rounds, indexed by $t = 0, \ldots, T-1$. Formally, the state of the system at iteration t is formulated as:

$$S^t = (P, I^t, \mathcal{A}, \mathcal{F}, \mathcal{R}^t, H^t). \tag{1}$$

Within each round t, the debaters, Moderator, and Regulator operate on the same current image I^t. Each agent $\mathcal{A}_i$ applies a function $f \in \mathcal{F}$ to engage in the debate and produces a response:

$$R_i^t = f(P, I^t, \mathcal{A}_i). \tag{2}$$

The replies at each iteration comprise three types of outputs:

$$R^t \in (D^t, M^t, U^t). \tag{3}$$

D^t is the set of debate replies at iteration t, M^t is the moderator output, and U^t is the regulator instruction.

In addition to R^t, the system maintains a persistent memory that spans the whole debate:

$$H^t = (C^t, E^t). \tag{4}$$

C^t is the running criteria set accumulated up to iteration t. E^t is the execution history accumulated up to iteration t.

3.2 Debate Dynamics Modeling and Control

The first challenge in MAD is the felicity gap of multimodal clarification. MAD differs from textual debate in that the debated topic is not static. The image state I^t changes over iterations, so the salient disagreements also change. To make this process tractable, we maintain an explicit criteria roadmap as part of the debate history.

To connect the roadmap to the debate flow, we instantiate two adversarial debaters, namely A_p and A_o. At iteration $t = 0$, the two debaters independently audit the initial image against the prompt and each produces an evidence-bearing debate claim:

$$D^0 = f_{\text{init}}(A_p, A_o, P, I^0). \tag{5}$$

Subsequent rounds use an asymmetric exchange between A_o and A_p, guided by the running criteria C^t:

$$D^t = f_{\text{debate}}(A_p, A_o, P, I^t, C^t), \qquad 1 \le t < T. \tag{6}$$

where D^t is the debate output at iteration t. The running set C^t acts as a refinement roadmap. It converts the sparse prompt into a list of verifiable checks that the debaters must follow.

Instead of relying on free-form impressions, the debate output $D^t = \{(d^t_{p,1}, d^t_{p,2}), (d^t_{o,1}, d^t_{o,2})\}$ is anchored to what can be confirmed from the current image. Each debater A_i, where $i \in \{p, o\}$, outputs its own pair $(d^t_{i,1}, d^t_{i,2})$ at iteration t. $d^t_{i,1}$ specifies the visual targets to be checked under the running criteria set C^t. $d^t_{i,2}$ reports the corresponding states. A_o raises disputed targets and unmet constraints. A_p defends the current image under the same criteria.

3.3 Regulator for Feasibility-Aware Control

The second challenge in MAD is that semantic fidelity does not imply execution feasibility. Debate agents may produce linguistically precise clarifications, yet the downstream executor may fail to instantiate them as concrete visual edits.

A feasibility-sensitive control mechanism is therefore required to preserve the normative force of debate while keeping the loop actionable. We address this gap by introducing a regulator agent A_r. A_r occupies the interface between language-based deliberation and visual execution.

At iteration t, regulator agent A_r receives the debate history $H^t = (C^t, E^t)$ and the moderator judgement M^t. Formally, the regulator decision is obtained by maximizing a probability π and can be equivalently written as a function in $\mathcal{F}$:

$$U^t = \arg\max_U \ \pi\big(U \mid A_r, P, I^t, H^t, M^t\big), \tag{7}$$

$$= f_{\text{regulate}}(A_r, P, I^t, H^t, M^t) = \{u^t\}. \tag{8}$$

Here u^t denotes the realized regulation at iteration t. It either authorizes an executable visual edit or adds a new criterion. The regulator selects between them to align debated intent with the executor's feasible action space, and to maintain a tractable refinement trajectory under the accumulated history.

The regulation u^t admits two realizations. If the realized regulation is an executable instruction, we denote it by $u^t = e^t$. The execution record is updated as

$$E^{t+1} = E^t + \big(1 - \mathbb{1}_{E^t}(u^t)\big)\, \delta(u^t, e^t)\, \delta_{u^t}, \tag{9}$$

where $\mathbb{1}_{E^t}(u^t)$ indicates whether a regulation u^t has been recorded in E^t (equal to 1 if recorded, and 0 otherwise). $\delta(u^t, e^t)$ is the Kronecker delta, where $\delta(u^t, e^t) = 1$ if $u^t = e^t$ and $\delta(u^t, e^t) = 0$ otherwise, so $\delta(u^t, e^t)$ activates the execution operation. δ_{u^t} denotes the unit record item indexed by u^t, which adds an entry at the current output without affecting other recorded entries. Then it is executed by the downstream editor as $I^{t+1} = f_{\text{edit}}(A_e, I^t, e^t)$.

If the realized regulation is a new criterion, we denote it by $u^t = c^t$. The criteria record is updated as

$$C^{t+1} = C^t + \big(1 - \mathbb{1}_{C^t}(u^t)\big)\, \delta(u^t, c^t)\, \delta_{u^t}. \tag{10}$$

Here $\mathbb{1}_{C^t}(u)$ is defined for the criteria record, and $\delta(u^t, c^t)$ activates the criteria-adding operation. A request for re-debate updates the criteria and consumes an outer round, while leaving the image and execution history unchanged: $I^{t+1} = I^t$ and $E^{t+1} = E^t$.

3.4 Moderator for Factual Termination

The third challenge in MAD concerns moderation. In textual debate, a judge can terminate deliberation by reflective judgement over linguistic content. In MAD, the target is factual and cross-modal. Termination must be based on whether the current image satisfies the prompt.

At iteration t, A_m is designed to evaluate the prompt P against the current image I^t under the accumulated debate history $H^t = (C^t, E^t)$. Formally, the moderator output is generated by a moderation function in $\mathcal{F}$:

$$M^t = f_{\text{moderate}}(A_m, P, I^t, D^t, H^t). \tag{11}$$

We formulate the moderator output as

$$M^t = (m_1^t,\ m_2^t), \tag{12}$$

where $m_1^t \in \{0, 1\}$ is a binary satisfaction judgement indicating whether the current image already satisfies the prompt. m_2^t is a set of verifiable diagnostic conditions derived from the consolidated debate.

MAD stops immediately when the Moderator judges the current image to satisfy the prompt and returns that image. Otherwise, the diagnostics are passed to the Regulator, which authorizes an edit or requests re-debate. If all T rounds are exhausted without satisfaction, MAD returns the final image I^T, without an additional Moderator evaluation:

$$I^* = \begin{cases} I^{t_s}, & \text{if satisfaction is first detected at round } t_s, \\ I^T, & \text{otherwise.} \end{cases} \tag{13}$$

We summarize the agents and functions used throughout the method. The instantiated agents form the set

$$A = \{A_p, A_o, A_m, A_r, A_g, A_e\}, \tag{14}$$

and the core functions that parameterize the debate and refinement loop form the set

$$\mathcal{F} = \{f_{\text{init}}, f_{\text{debate}}, f_{\text{regulate}}, f_{\text{moderate}}, f_{\text{edit}}\}. \tag{15}$$

Here A_p and A_o are the debaters, A_m is the moderator, A_r is the regulator, and A_g and A_e are the generator and editor. The functions f_{init}, f_{debate}, f_{moderate}, f_{regulate}, and f_{edit} respectively initialize the debate, update claims under evolving criteria, produce factual moderation, map diagnostics to feasible actions, and apply the selected edit to the next image I^{t+1}.

4 Experiments

4.1 Experimental setup

Benchmarks and models. MAD is evaluated on GenEval Ghosh et al. (2023), T2I-CompBench Huang et al. (2023), TIIF-Bench Wei et al. (2025), and COCO-5K. Qwen3-VL-8B Bai et al. (2025) is used for debate, moderation, and regulation. Results with InternVL backbones are reported in the Appendix. MAD is applied to frozen generator–editor pairs, including SDXL Podell et al. (2023), SD3 Esser et al. (2024), Kontext Labs et al. (2025), and FLUX.2 Labs (2025). Prompt templates are provided in Appendix C.

Protocol. Unless stated otherwise, we run at most T=2 debate rounds. Within each Base/MAD pair the generator, sampling, and evaluator are held fixed. Full category tables and evaluation protocols are in Appendix A.

Paired comparisons. We conduct performance comparisons on GenEval, T2I-CompBench, and COCO-5K. Paired results are reported for pretrained and reasoning-oriented generation pipelines before and after MAD refinement.

4.2 Performance Comparison to T2I Methods

Performance on Image Quality and Alignment Metrics. Table 1 reports image quality and alignment metrics on GenEval, T2I-CompBench, and COCO-5K. The most consistent effect is on distributional alignment, since CLIP-FID is reduced in every reported Base/MAD pair across all three benchmarks. The complementary gains in IS and CLIPScore are frequent. TwiG improves all three metrics on all three benchmarks, and Kontext obtains the largest GenEval gains across the three metrics. On COCO-5K, CLIPScore is improved for every base system, and all three metrics are improved for five of the six systems. Occasional regressions remain, such as the IS decrease for SDXL. We additionally evaluate MAD on TIIF-Bench with a VLM-based protocol, and the detailed results are reported in Appendix A.

Method	Paradigm	GenEval			T2I-CompBench			COCO-5K		
		CLIP-FID ↓	IS ↑	CLIPScore ↑	CLIP-FID ↓	IS ↑	CLIPScore ↑	CLIP-FID ↓	IS ↑	CLIPScore ↑
Base/MAD Comparisons										
Kontext Labs et al. (2025)	Pretrain	37.13	25.48	28.49	31.03	**30.84**	27.29	16.44	30.87	26.51
Kontext+**MAD**	Training-free	**32.59** ↓4.54	**27.71** ↑2.23	**29.60** ↑1.11	**30.28** ↓0.75	29.78 ↓1.06	**27.55** ↑0.27	**16.11** ↓0.33	**31.92** ↑1.04	**26.62** ↑0.11
SDXL Podell et al. (2023)	Pretrain	35.92	32.69	29.23	30.84	**37.11**	28.10	14.50	**42.12**	27.66
SDXL+**MAD**	Training-free	**34.49** ↓1.43	**32.74** ↑0.05	**29.45** ↑0.22	**29.79** ↓1.06	35.34 ↓1.76	**28.17** ↑0.07	**13.87** ↓0.63	41.99 ↓0.13	**27.71** ↑0.05
ReasonGen-R1 Zhang et al. (2025b)	SFT+RL	39.75	22.61	29.96	30.53	28.73	**29.18**	16.50	39.13	27.79
ReasonGen-R1+**MAD**	Training-free	**37.98** ↓1.77	**22.96** ↑0.35	**30.05** ↑0.08	**29.14** ↓1.39	**28.78** ↑0.05	28.90 ↓0.28	**14.87** ↓1.63	**39.40** ↑0.27	**27.81** ↑0.02
GoT Fang et al. (2025)	Pretrain+SFT	45.35	**15.86**	28.90	36.21	**27.31**	26.93	21.10	30.27	26.31
GoT+**MAD**	Training-free	**43.22** ↓2.13	15.13 ↓0.73	**29.11** ↑0.21	**34.26** ↓1.95	26.65 ↓0.66	**27.25** ↑0.32	**19.35** ↓1.75	**30.34** ↑0.07	**26.55** ↑0.24
MILR Mi et al. (2026)	Training-free	37.04	25.47	**30.11**	29.59	26.22	27.60	14.62	34.74	27.70
MILR+**MAD**	Training-free	**35.12** ↓1.92	**25.90** ↑0.43	30.10 ↓0.01	**27.56** ↓2.03	**26.97** ↑0.75	**28.05** ↑0.45	**13.38** ↓1.24	**35.94** ↑1.20	**27.82** ↑0.12
TwiG Guo et al. (2026)	Training-free	35.83	26.74	28.04	30.11	28.22	27.28	16.49	30.50	26.80
TwiG+**MAD**	Training-free	**33.58** ↓2.25	**26.87** ↑0.13	**28.89** ↑0.86	**28.42** ↓1.69	**29.76** ↑1.55	**27.74** ↑0.45	**14.63** ↓1.86	**31.65** ↑1.15	**27.28** ↑0.48

Table 1: Performance on GenEval, T2I-CompBench, and COCO-5K. Bold indicates the better result within each Base/MAD pair, and arrows report absolute changes in the original metric units.

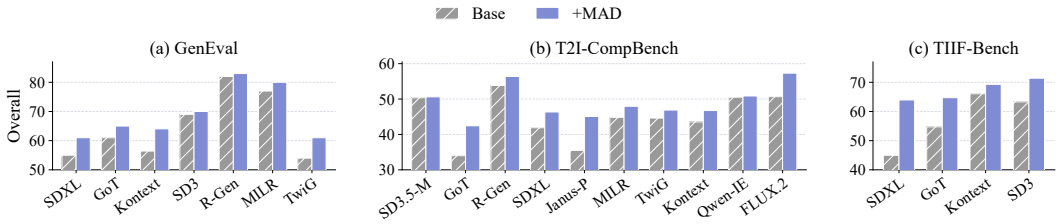

Figure 4: Overall scores of Base versus MAD on GenEval, T2I-CompBench, and TIIF-Bench. T2I-CompBench Overall is the unweighted mean of six category scores.

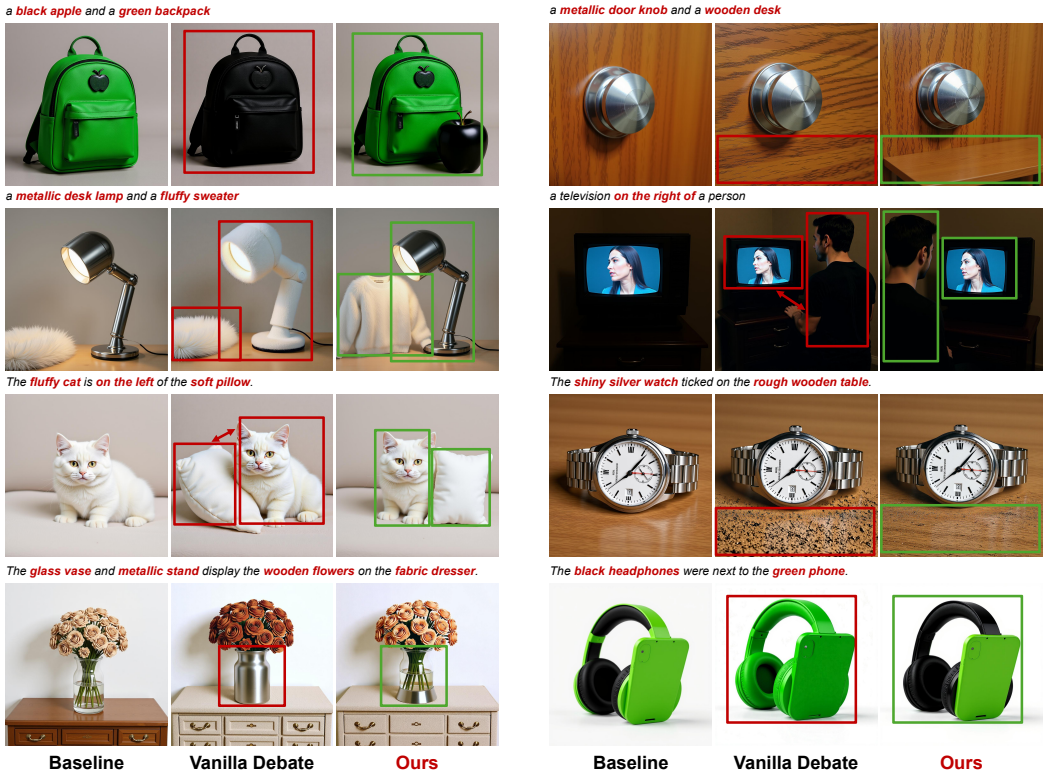

Figure 5: Qualitative T2I-CompBench comparisons among Baseline, Vanilla Debate, and MAD (Ours). Each triplet uses the same prompt.

Performance on Compositional Benchmarks. Fig. 4 summarizes Overall scores on GenEval, T2I-CompBench, and TIIF-Bench. MAD improves every reported Base/MAD pair across the three benchmarks. These gains span generators with substantially different baseline Overall scores, supporting MAD's generalizability across the evaluated generator families. Category-level results for

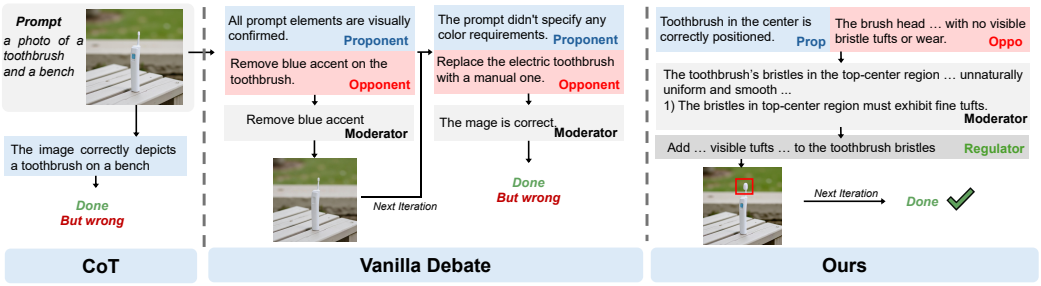

Figure 6: Process-level comparison of CoT, Vanilla Debate, and MAD (Ours) on the same prompt.

Method	Components				Overall ↑	Category scores (↑)					
	Deb.	Crit.	Fact.	Reg.		Color	Shape	Texture	Spatial	Non-sp.	Complex
Baseline	–	–	–	–	0.4353	0.7237	0.4713	0.5922	0.2725	0.2604	0.2918
Single Critic	–	–	–	–	0.4504 ↑1.5	0.7201	0.5079	0.6149	0.3040	0.2406	0.3148
Vanilla Debate	✓	–	–	–	0.4489 ↑1.4	0.6988	0.4992	0.6281	0.2890	0.2594	0.3191
w/ Regulator	✓	–	–	✓	0.4434 ↑0.8	0.6902	0.4878	0.6240	0.2762	0.2606	0.3216
w/ Criteria	✓	✓	–	–	0.4614 ↑2.6	0.7462	0.4939	0.6518	0.3045	0.2602	0.3117
w/ Criteria and Factual	✓	✓	✓	–	0.4586 ↑2.3	0.7312	0.5151	0.6270	0.2977	0.2595	0.3209
MAD w/ Single-Critic Reasoning	–	✓	✓	✓	0.4550 ↑2.0	0.7258	0.5126	0.6287	0.2985	0.2563	0.3081
MAD w/ Static Criteria	✓	*sta*	✓	✓	0.4624 ↑2.7	0.7256	0.5147	0.6597	0.2946	0.2600	0.3195
MAD w/ Reflective Moderator	✓	✓	*ref*	✓	0.4417 ↑0.6	0.7215	0.4802	0.6146	0.2788	0.2600	0.2953
MAD w/ Direct Execution	✓	✓	✓	–	0.4616 ↑2.6	0.7607	0.5091	0.6326	0.2962	0.2607	0.3100
MAD (Ours)	✓	✓	✓	✓	**0.4675** ↑3.2	0.7616	0.5184	0.6455	0.3076	0.2608	0.3109

Table 2: Ablation on T2I-CompBench. Overall is the unweighted mean of six categories. Arrows report percentage-point changes vs. Baseline. *sta* denotes static criteria and *ref* denotes reflective moderation.

GenEval and T2I-CompBench are provided in Appendix A. The full TIIF-Bench comparison and an efficiency comparison with iterative T2I methods are also reported there.

4.3 Qualitative Analysis

Fig. 5 and Fig. 6 compare MAD with Baseline, Vanilla Debate, and CoT. The T2I-CompBench examples cover color, spatial relations, texture, and complex composition. Baseline and Vanilla Debate exhibit object fusion, incorrect attribute binding, and missing relations or materials in these examples. MAD preserves the specified objects and their required attributes and relations. Fig. 6 illustrates the refinement process for the prompt "a photo of a toothbrush and a bench." CoT uses the debate model as a single-agent reasoner and accepts the initial image. Vanilla Debate discusses color and device type, which are not specified in the prompt. Both terminate without correcting the missing bristle texture. MAD identifies this discrepancy and specifies a verifiable success criterion. An executable edit is then produced to correct the image. Additional examples are provided in Appendix B.

4.4 Ablation Study

Table 2 compares full MAD with refinement baselines and controlled variants. Baseline generates images without refinement. Single Critic directly proposes edits from image–text discrepancies. Vanilla Debate introduces proponent–opponent exchanges and a moderator that produces edit instructions. Its additive variants introduce the Regulator, dynamic criteria, or dynamic criteria with factual moderation.

MAD w/ Single-Critic Reasoning replaces the two debaters with a single self-refining critic under a matched computational budget. All other components are retained. The remaining controlled variants separately simplify dynamic criteria, factual moderation, or regulated execution. Detailed definitions are provided in Appendix B.

Component contributions. Full MAD achieves the highest Overall and improves all six categories over Baseline. Under a matched computational budget, it outperforms MAD w/ Single-Critic Reasoning in every category, supporting the contribution of adversarial debate. Dynamic criteria improve Vanilla Debate. Within full MAD, fixing the criteria, removing the Moderator's image input, or replacing the Regulator reduces Overall. Removing the Moderator's image input causes the largest decrease, supporting direct visual verification. MAD leads Color, Shape, Spatial, and Non-spatial, while MAD w/ Static Criteria leads Texture and Vanilla Debate w/ Regulator leads Complex. The largest gains over MAD w/ Single-Critic Reasoning are observed in Color and Texture, with improvements also observed in both relation categories. Full MAD outperforms w/ Criteria and Factual in five of the six categories. These results support the combined use of adversarial interaction and regulated execution alongside dynamic criteria and factual moderation.

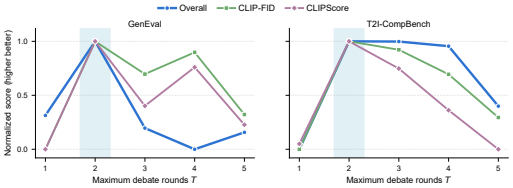

Figure 7: GenEval and T2I-CompBench versus maximum debate rounds T.

Maximum debate rounds. Fig. 7 varies T from 1 to 5 on GenEval and T2I-CompBench. Overall, CLIP-FID, and CLIPScore peak at T=2. Larger T is not consistently better. This supports the default T=2 in the main experiments.

Executability and repetition of visual execution. Fig. 8 reports diagnostics for visual edit instructions e^t. ***Executability*** measures whether an instruction specifies concrete visual edits that can be performed by the executor. ***Repetition*** identifies near-identical instructions across debate rounds. On GenEval, executable instructions account for 74.6%, 76.2%, and 81.0% with Single Critic, Vanilla Debate, and MAD, respectively. The corresponding values on T2I-CompBench are 65.8%, 74.5%, and 75.2%. Repetitive instructions decrease from 42.2% with Vanilla Debate to 10.9% with MAD on GenEval and from 39.7% to 10.5% on T2I-CompBench.

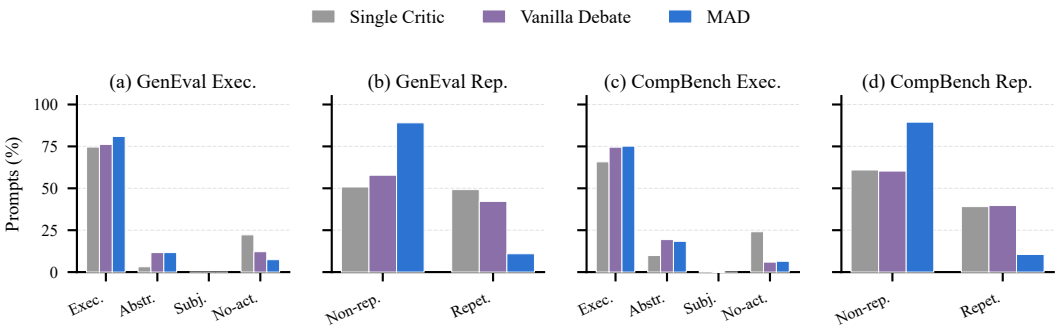

Figure 8: Instruction e^t executability and repetition for Single Critic, Vanilla Debate, and MAD on GenEval and T2I-CompBench. Each diagnostic is reported as a percentage of eligible prompts.

Further ablations. Additional results on TIIF-Bench, backbones, and GenEval components are in Appendix B. On TIIF-Bench, full MAD achieves the highest Overall, Basic Following, and Advanced Following scores among the ablated variants. On GenEval, it also achieves the highest Overall score, although its advantage over Single Critic is small.

5 Conclusion

This work addresses the gap between linguistic intent and perceptual evidence in iterative text-to-image generation. We propose MAD, a vision-grounded adversarial deliberation framework. It uses an explicit evolving criterion set, evidence-grounded claims, and moderator-produced verifiable diagnostics with stopping conditions. A history-aware Regulator then selects between executable edits and criterion-injected re-debate. Across GenEval, T2I-CompBench, and TIIF-Bench, MAD is a training-free plug-in that improves Overall, with model-dependent category-level trade-offs.

AI ASSISTANCE

Generative AI tools were used to assist with drafting, translation, language editing, figure preparation, and manuscript formatting.

REFERENCES

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19396–19405, 2025a.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025b.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

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

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Hao Tian, Shilin Yan, Weihao Yu, Xingyu Zeng, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability of MLLM for visual generation and editing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=eCElREEUsr.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

Ziyu Guo, Renrui Zhang, Hongyu Li, Manyuan Zhang, Xinyan Chen, Sifan Wang, Yan Feng, Peng Pei, and Pheng-Ann Heng. Thinking-while-generating: Interleaving textual reasoning throughout visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26295–26305, 2026.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15839–15853, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.770. URL https://aclanthology.org/2025.acl-long.770/.

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.

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, 2025. URL https://arxiv.org/abs/2505.00703.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13263–13272, June 2025.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding, May 2024.

Dayong Liang, Xiao-Yong Wei, and Changmeng Zheng. Multi-agent undercover gaming: Hallucination removal via counterfactual test for multimodal reasoning. *arXiv preprint arXiv:2511.11182*, 2025.

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

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

Yapeng Mi, Yanpeng Zhao, Henry Li, Chenxi Li, Huimin Wu, Xiaojian Ma, Song-Chun Zhu, Yingnian Wu, and Qing Li. Milr: Improving multimodal image generation via test-time latent reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 21661–21681, 2026.

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.

Haonan Sheng, Dou Hu, Lingwei Wei, Wei Zhou, and Songlin Hu. Regularized contrastive decoding with hard negative samples for llm hallucination mitigation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6061–6073, 2025.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xiao-Yong Wei and Zhen-Qun Yang. Coached active learning for interactive video search. In *Proceedings of the 19th ACM international conference on Multimedia*, 2011.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

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

Zhen-Qun Yang Xiao-Yong Wei. Coaching the exploration and exploitation in active learning for interactive video retrieval. *IEEE Transactions on Image Processing*, 22(3):955–968, 2013.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer, March 2025.

Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Ziyu Guo, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Peng Gao, and Hongsheng Li. Let's verify and reinforce image generation step by step. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28662–28672, 2025a.

Xulu Zhang, Xiao-Yong Wei, Jinlin Wu, Tianyi Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Compositional inversion for stable diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7350–7358, 2024.

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

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 419–428, 2024.