LEARNING FROM REFLECTION: ON-POLICY SELF-DISTILLATION FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models can diagnose errors in their own generations, providing image-level feedback for text-to-image post-training. However, existing methods typically use this feedback as sparse rewards to optimize the model, making it difficult to assign credit to individual generation states across timesteps. We propose Text-to-Image On-Policy Self-Distillation (T2I-OPSD), which turns textual reflections into dense state-level supervision, providing explicit velocity correction targets to improve prompt alignment. Specifically, the current model acts as both student and self-teacher, with the latter additionally conditioned on the student's generated image and textual reflection. By matching their velocity predictions at shared student-induced states, we turn image-specific feedback into dense directional supervision that shapes the student's velocity field for subsequent prompt-only generation. Furthermore, to stabilize supervision as the student evolves, we use reference-centered residual distillation to interpolate self-teacher predictions with those of a frozen prompt-only reference. We evaluate T2I-OPSD on multiple text-to-image benchmarks, achieving leading performance in compositional generation without additional sampling steps or inference memory.

1 INTRODUCTION

Text-to-image generation has developed along two broad directions: dedicated diffusion and flow-based generators (Ho et al., 2020; Lipman et al., 2022; Liu et al., 2022; Esser et al., 2024; Black Forest Labs, 2024), and unified multimodal models (UMMs) (Deng et al., 2025; Team, 2024; Wang et al., 2024; Chen et al., 2025b; Xie et al., 2025). By integrating visual understanding and image generation, UMMs enable the same model to synthesize an image, inspect it, and reason about how to correct it. This integration creates an opportunity to turn a model's understanding of its own outputs into feedback for improving generation.

Leveraging their visual understanding and reasoning capabilities, UMMs support semantic planning before generation (Jiang et al., 2025; Duan et al., 2025) and reflective refinement afterward (Huang et al., 2026a;b). For post-training, most existing methods use self-evaluation to construct rewards from image alignment, object-level quality, or verifiable answers (Mao et al., 2025; Jin et al., 2025; Hong et al., 2026; Pan et al., 2026; Huang et al., 2026a). However, these rewards provide sparse feedback on completed images, making it difficult to assign credit to individual generation states across timesteps. The key challenge is therefore to translate reflections on generation errors into dense state-level supervision. By providing explicit velocity correction targets across noise levels, such supervision guides how predictions should change at individual states and enables the model to internalize corrective feedback into its parameters for subsequent prompt-only generation.

Therefore, we introduce Text-to-Image On-Policy Self-Distillation (T2I-OPSD), which turns reflective feedback into dense state-level velocity supervision across timesteps to improve prompt alignment in prompt-only generation. Given a prompt c, the current student generates an image $\hat{x}_0$ and then produces a textual reflection r on it. The image and reflection $(\hat{x}_0, r)$ serve as privileged context for a self-teacher that shares the student's current parameters. Using this context together with the prompt c, the self-teacher predicts velocity at the same student-induced query state x_t as the prompt-only student. We then train the student to match these reflection-conditioned predictions, internalizing the corrective feedback in its velocity field for subsequent prompt-only generation.

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.



Figure 1: **Performance overview.** T2I-OPSD improves generation quality and prompt alignment without additional inference overhead. $1\times$ and $2\times$ denote the number of image-generation rounds at inference. +SRR denotes AlphaGRPO with self-reflective refinement.

As training progresses, the student's generation distribution evolves, changing the images and reflections that constitute the self-teacher's privileged context, which we term endogenous privilege. The reflection capability may not generalize to this evolving distribution, resulting in unreliable teacher corrections. Continued distillation from these corrections can degrade the student, whose subsequent generations further weaken the supervision, creating a feedback loop of coupled degradation that we call co-drift. To mitigate this instability, we introduce reference-centered residual distillation, which anchors the supervision target to a frozen reference. We form the target by interpolating between the predictions of the frozen reference and the reflection-conditioned self-teacher, retaining reflective corrections while limiting deviation from the reference prediction. T2I-OPSD retains only the prompt-only student at deployment and requires no additional sampling steps or inference memory.

We evaluate T2I-OPSD on multiple text-to-image benchmarks, achieving leading performance in compositional generation. Notably, T2I-OPSD outperforms AlphaGRPO (Huang et al., 2026a) with self-reflective refinement (SRR) even under prompt-only conditioning. Specifically, T2I-OPSD improves GenEval (Ghosh et al., 2023) from 76.77 for our reproduced BAGEL baseline to 85.04 and achieves 88.77 on DPG-Bench (Hu et al., 2024), surpassing AlphaGRPO (86.05) and AlphaGRPO+SRR (85.90).

Our contributions are threefold:

- We introduce T2I-OPSD, which converts predictions conditioned on a self-generated image and reflection into dense velocity supervision at shared student-induced query states, internalizing reflective feedback into model parameters to improve prompt-only generation.
- We formalize endogenous privilege and the resulting co-drift, revealing a feedback loop between student degradation and unreliable supervision. To address this, we introduce reference-centered residual distillation to stabilize training.
- We demonstrate that T2I-OPSD achieves leading performance across multiple text-to-image benchmarks without additional sampling steps or inference memory.

2 Related Work

2.1 Post-Training for Text-to-Image Generation

Post-training improves prompt alignment in diffusion and flow-based generators (Esser et al., 2024; Black Forest Labs, 2024) through reward and preference optimization (Black et al., 2024; Fan et al., 2023; Wallace et al., 2024; Liu et al., 2025; Xue et al., 2025; Sun et al., 2026; Zheng et al., 2026). Unified multimodal models (UMMs) (Team, 2024; Wang et al., 2024; Chen et al., 2025b; Zhou et al., 2025; Xie et al., 2025; Deng et al., 2025; Pan et al., 2025; Wu et al., 2026a) extend this paradigm

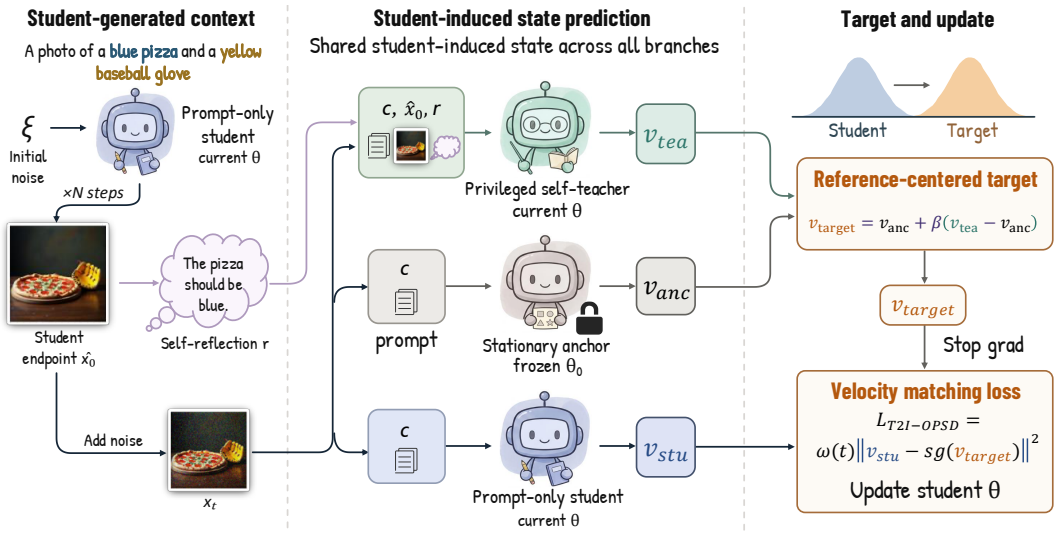


Figure 2: **Overview of T2I-OPSD.** The prompt-only student generates an endpoint image, and the same model then produces a textual reflection on it. Using the image and reflection as privileged context, the self-teacher predicts velocities at the same student-induced states as the student. We then interpolate these predictions with those of a frozen reference to form the supervision target.

to generation planning, joint understanding-generation alignment, and self-evaluation (Jiang et al., 2025; Duan et al., 2025; Mo & Yun, 2026; Mao et al., 2025; Jin et al., 2025; Hong et al., 2026; Pan et al., 2026; Thawkar et al., 2026). AlphaGRPO and R3 leverage textual reflection for regeneration and editing (Huang et al., 2026a; Ye et al., 2026), whereas RA-GRPO internalizes numerical trajectory corrections through counterfactual local supervision (Wu et al., 2026b). REST reuses reward-scored teacher trajectories for few-step distillation (Li et al., 2026b). T2I-OPSD instead uses textual reflection as teacher-side conditioning for local velocity supervision.

2.2 On-Policy (Self-)Distillation

On-policy distillation trains on student-generated states, while privileged-context self-distillation creates teacher advantage through additional conditioning (Zhao et al., 2026). DiffusionOPD matches local teacher transitions along student trajectories (Li et al., 2026a), while Flow-OPD integrates multiple experts with manifold-anchor regularization (Fang et al., 2026). STEP-OPD further extrapolates teacher velocities relative to a frozen base and aligns the direction and magnitude of internal representation changes (Wei et al., 2026). D-OPSD uses paired target images as privileged context to continually tune few-step models through same-model distillation on student trajectories (Jiang et al., 2026), whereas Any-OPD connects heterogeneous models through output-representation matching (Fu et al., 2026). DiffusionOPSD constructs detached intermediate clean-output targets using reward gradients around behavior-policy anchors (Zhou et al., 2026). T2I-OPSD instead matches reflection-conditioned velocities from a same-parameter self-teacher at shared student-induced query states, using the student's own image and reflection.

3 Methods

We introduce Text-to-Image On-Policy Self-Distillation (T2I-OPSD), a framework that converts a UMM's reflection-conditioned predictions into dense velocity supervision for prompt-only text-to-image generation. First, we instantiate the method from a UMM capable of diagnosing generation errors and revising its output accordingly (Section 3.1). Next, we condition the self-teacher on privileged image and reflection context and distill its velocity predictions into the prompt-only student at shared student-induced states (Section 3.2). We then analyze endogenous privilege and the resulting co-drift, which can undermine teacher reliability and destabilize training (Section 3.3). Finally,

we introduce reference-centered residual distillation to stabilize training by interpolating between self-teacher and frozen prompt-only reference predictions (Section 3.4).

3.1 Reflection as a Source of Privileged Supervision

Our method builds on a UMM with self-reflective refinement (SRR) capabilities, which enable it to diagnose discrepancies between a generated image and the prompt through textual reflection and revise the image accordingly. Prior work has equipped BAGEL (Deng et al., 2025) with these capabilities through reinforcement learning (Huang et al., 2026a) and supervised fine-tuning (Huang et al., 2026b) to correct attribute and spatial-relation errors and improve prompt alignment. We use the generated image and its reflection as privileged context to guide the self-teacher's velocity predictions, providing corrective targets for the prompt-only student.

T2I-OPSD does not elicit SRR from scratch, but directly uses the publicly released AlphaGRPO checkpoint (Huang et al., 2026a) as a reflection-capable initialization.

3.2 Student-Induced State Velocity Distillation

Forward-Resampled Transition Matching T2I-OPSD matches local transitions under two conditioning contexts at shared student-induced states. Conditioned on a prompt $c \sim \mathcal{D}$, the prompt-only student generates an endpoint image $\hat{x}_0$ from Gaussian noise ξ, and the same model then produces a textual reflection r on this image.

$$\xi \sim \mathcal{N}(0, I), \quad \hat{x}_0 = \mathrm{Sample}_\theta(\xi; c), \quad r = \mathcal{R}_\theta(c, \hat{x}_0). \tag{1}$$

Let $q_\theta(\cdot \mid c)$ denote the resulting endpoint distribution, with student and teacher contexts defined as $c_S = c$ and $c_T = (c, \hat{x}_0, r)$, respectively. To compare their local transitions at a shared student-induced state, we forward-noise the current student endpoint as $x_t = \alpha_t \hat{x}_0 + \sigma_t \epsilon$. At this state, $p_\theta^\Delta(\cdot \mid x_t, t, c_B)$ denotes the local transition kernel from t to $t - \Delta$ under context $c_B \in \{c_S, c_T\}$. Matching the transitions under these two contexts gives the objective

$$\mathcal{L}_{\text{T2I-OPSD}}^{\text{KL}} = \mathbb{E}_{\substack{c\sim\mathcal{D},\, \hat{x}_0\sim q_\theta(\cdot|c),\\ t\sim p(t),\, \epsilon\sim\mathcal{N}(0,I)}} \left[\mathrm{KL}\left(p_\theta^\Delta(\cdot \mid x_t, t, c_S) \,\middle\|\, p_\theta^\Delta(\cdot \mid x_t, t, c_T) \right) \right]. \tag{2}$$

For rectified flow, $\alpha_t = 1 - t$ and $\sigma_t = t$. The two transition kernels in Eq. 2 use the same parameters θ and differ only in their conditioning contexts. During optimization, the transition conditioned on c_T is treated as a stop-gradient target.

Forward resampling requires only the endpoint $\hat{x}_0$, avoids caching the reverse trajectory, and decouples endpoint generation from distillation-state construction, making the method compatible with different ODE, SDE, or accelerated samplers. Furthermore, AlphaGRPO provides more reliable reflective corrections when regenerating images with independently sampled noise, as detailed in App. D. Therefore, we draw ϵ independently of the initial sampling noise ξ so that the student-induced state is not tied to the original rollout seed.

From Transition KL to Velocity Distillation Under a local Gaussian transition parameterization, a velocity prediction v_B under context c_B induces the one-step transition

$$p_\theta^\Delta(\cdot \mid x_t, t, c_B) = \mathcal{N}\left(\cdot; a_t x_t + b_t v_B, \tau_t^2 I\right), \qquad (c_B, v_B) \in \{(c_S, v_{\text{stu}}), (c_T, v_{\text{tea}})\}. \tag{3}$$

Here, a_t and b_t are shared transition coefficients, and $\tau_t > 0$ is the shared transition standard deviation, with their dependence on Δ suppressed. The student and teacher velocities are

$$v_{\text{stu}} = v_\theta(x_t, t \mid c_S), \qquad v_{\text{tea}} = v_\theta(x_t, t \mid c_T). \tag{4}$$

Because the two conditional transitions share the same state, noise schedule, and covariance, their local KL depends only on the difference between the transition means:

$$\mathrm{KL}\left(p_\theta^\Delta(\cdot \mid x_t, t, c_S) \,\middle\|\, p_\theta^\Delta(\cdot \mid x_t, t, c_T) \right) = \underbrace{\frac{b_t^2}{2\tau_t^2}}_{\omega(t)} \left\| v_{\text{stu}} - v_{\text{tea}} \right\|^2. \tag{5}$$

The transition objective in Eq. 2 therefore reduces to the trainable student-induced state velocity-distillation loss

$$\mathcal{L}_{\text{OPSD}} = \mathbb{E}_{\substack{c\sim\mathcal{D},\,\hat{x}_0\sim q_\theta(\cdot|c),\\ t\sim p(t),\,\epsilon\sim\mathcal{N}(0,I)}} \left[\omega(t)\left\|v_\theta(x_t,t\mid c_S) - \text{sg}\left[v_\theta(x_t,t\mid c_T)\right]\right\|^2\right]. \tag{6}$$

Forward resampling determines where distillation is applied, while the reflection-conditioned branch provides a candidate corrective velocity at each shared student-induced state. Sampling t across noise levels turns reflection on completed images into dense state-level supervision across timesteps. Compared with supervised fine-tuning on fixed image datasets, online distillation keeps corrective supervision aligned with the student's evolving generation distribution and focuses learning on errors in its current outputs.

3.3 Endogenous Privilege and Co-Drift

Exogenous Supervision and Fixed Criteria Many reinforcement learning and OPD/OPSD methods define supervision using external rewards, fixed references, or independently trained teachers (Xue et al., 2025; Liu et al., 2025; Zheng et al., 2026; Li et al., 2026a; Fang et al., 2026; Jiang et al., 2026). As the student distribution evolves, this external anchor maintains a consistent supervision criterion, supporting stable optimization. We refer to this supervision as *exogenous privilege*, as illustrated in Fig. 3 (a).

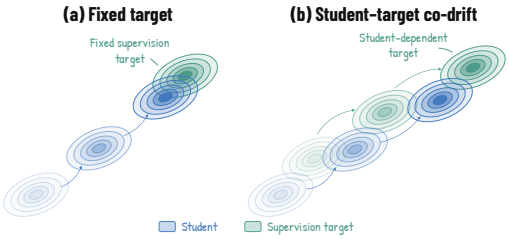


Figure 3: Independently trained teachers define student-independent criteria. T2I-OPSD's privileged context evolves with the student distribution, creating a feedback loop.

Endogenous Privilege and Target Refresh In T2I-OPSD, the current student generates the image and reflection that form the self-teacher's privileged context. At optimization step k, the parameters θ_k determine these inputs and the induced state. To track their changes, we hold c, ξ, t, and ϵ fixed across adjacent steps and write

$$\hat{x}_{0,k} = \text{Sample}_{\theta_k}(\xi; c), \qquad x_{t,k} = \alpha_t \hat{x}_{0,k} + \sigma_t \epsilon, \qquad r_k = \mathcal{R}_{\theta_k}(c, \hat{x}_{0,k}). \tag{7}$$

We collect these dependencies in the self-teacher function $\mathcal{F}_\theta$ and denote the target at step k by T_k:

$$\mathcal{F}_\theta\left(\hat{x}_0; c, t, \epsilon\right) = v_\theta\left(\alpha_t \hat{x}_0 + \sigma_t \epsilon, t \mid c, \hat{x}_0, \mathcal{R}_\theta(c, \hat{x}_0)\right), \qquad T_k = \mathcal{F}_{\theta_k}\left(\hat{x}_{0,k}; c, t, \epsilon\right). \tag{8}$$

During each optimization step, the stop-gradient operation in Eq. 6 holds T_k fixed. After the student update, both the parameters and student-generated inputs change. By adding and subtracting $\mathcal{F}_{\theta_k}(\hat{x}_{0,k+1}; c, t, \epsilon)$, we decompose the target change exactly as

$$\begin{aligned} T_{k+1} - T_k = &\underbrace{\mathcal{F}_{\theta_{k+1}}\left(\hat{x}_{0,k+1}; c, t, \epsilon\right) - \mathcal{F}_{\theta_k}\left(\hat{x}_{0,k+1}; c, t, \epsilon\right)}_{\text{shared-parameter refresh}} \\ &+ \underbrace{\mathcal{F}_{\theta_k}\left(\hat{x}_{0,k+1}; c, t, \epsilon\right) - \mathcal{F}_{\theta_k}\left(\hat{x}_{0,k}; c, t, \epsilon\right)}_{\text{student-endpoint-induced target refresh}}. \end{aligned} \tag{9}$$

The first term captures the effect of shared-parameter updates with the new image held fixed. The second holds θ_k fixed and captures changes in the state, privileged image, and reflection induced by the new student output. Thus, the self-teacher's velocity target changes with the student's generations even at fixed parameters. We call this joint dependence of the supervised state and privileged context on the student *endogenous privilege*.

Positive Feedback and Co-Drift Endogenous privilege creates a cross-step feedback loop. A student update changes its generation distribution, which refreshes $\hat{x}_{0,k}$, r_k, and $x_{t,k}$ in subsequent iterations. The refreshed state and context then change the teacher target, and the updated target drives the next student update. Repeated target refreshes can therefore couple the evolution of the

student with that of its supervision signal, as illustrated in Fig. 3 (b). We refer to this coupled long-horizon degradation as *co-drift*. Stop-gradient blocks within-step gradient flow through the target but does not break this cross-step feedback loop.

Co-drift induces a progressive failure process. As the student distribution evolves, its generated images may gradually move beyond the regime in which the inherited SRR capability remains reliable. The resulting reflections and corrections can therefore become increasingly inaccurate, causing the privileged target to lose its advantage over prompt-only conditioning. Continuing to distill from such unreliable targets further shifts the student distribution and eventually destabilizes training.

3.4 Reference-Centered Residual Distillation

To stabilize supervision as the privileged context evolves, we anchor the target to the model's initial generation behavior. We leverage a frozen copy of the reflection-capable initialization θ_0 as a prompt-only reference. At the current student-induced state x_t, its velocity is

$$v_{\text{anc}}(x_t, t \mid c) = v_{\theta_0}(x_t, t \mid c). \tag{10}$$

We combine the reflection-conditioned self-teacher with the frozen prompt-only reference to construct the final supervision target:

$$v_{\text{target}} = v_{\text{anc}} + \beta\,(v_{\text{tea}} - v_{\text{anc}})\,, \qquad \beta \in [0, 1]. \tag{11}$$

Here, $v_{\text{tea}} = v_\theta(x_t, t \mid c, \hat{x}_0, r)$ is the reflection-conditioned self-teacher prediction, and β balances the self-teacher correction against the frozen prompt-only reference. The reference pulls the target toward the initial prompt-only dynamics, slowing distributional drift, preventing rapid divergence, and sustaining a more stable optimization plateau.

Training Objective. The velocity matching objective in Eq. 6 motivates a time-dependent weighting $\omega(t)$. For practical optimization, we adapt the self-normalized x_0-regression objective of DiffusionNFT (Zheng et al., 2026) to velocity distillation. Rather than explicitly specifying $\omega(t)$, we convert both the student velocity and the combined target velocity into endpoint predictions. Under the rectified-flow schedule, a velocity prediction v induces the corresponding x_0 prediction $x_t - tv$. During optimization, we keep the sampled states fixed and stop gradients through the supervision target. We then define

$$\begin{aligned} x_\theta(x_t, c, t) &= x_t - t\, v_\theta(x_t, t \mid c), \\ x_{\text{target}}(x_t, c, t) &= \text{sg}\left\{x_t - t\left[v_{\text{anc}} + \beta\,(v_{\text{tea}} - v_{\text{anc}})\right]\right\}. \end{aligned} \tag{12}$$

The final T2I-OPSD objective uses self-normalized endpoint-prediction regression:

$$\mathcal{L}_{\text{T2I-OPSD}} = \mathbb{E}_{\substack{c\sim\mathcal{D},\, \hat{x}_0\sim q_\theta(\cdot|c),\\ t\sim p(t),\, \epsilon\sim\mathcal{N}(0,I)}} \left[\frac{\|x_\theta(x_t, c, t) - x_{\text{target}}(x_t, c, t)\|_2^2}{\text{sg}\left[\text{mean}\left(|x_\theta(x_t, c, t) - x_{\text{target}}(x_t, c, t)|\right)\right]}\right]. \tag{13}$$

Under rectified flow, $x_\theta(x_t, c, t) - x_{\text{target}}(x_t, c, t)$ equals $-t\,(v_{\text{stu}} - v_{\text{target}})$ up to the stop-gradient operation on the target. The velocity-to-endpoint transformation therefore introduces timestep dependence through the factor t, and the resulting objective regresses the prompt-only student toward the reference-centered target at student-induced states.

Inference. After training, only the prompt-only student is retained for inference. Inference requires neither textual reflection nor a second image generation. Therefore, T2I-OPSD confines reflective correction to training and preserves the sampling steps, number of forward passes, and memory requirements of the original prompt-only model at deployment.

4 Experiments

4.1 Experimental Setup

We implement T2I-OPSD on BAGEL-7B-MoT (Deng et al., 2025), initialized from an AlphaGRPO-trained checkpoint (Huang et al., 2026a) with self-reflective refinement (SRR) capability. The student and self-teacher are instantiated from the same current parameters θ and differ only in conditioning context, while the anchor uses parameters θ_0 frozen at the start of training. We fix the target

Table 1: **Results on text-to-image benchmarks.** S and L denote short and long prompts on TIIF-Bench, respectively. IR / PS / HPS / Aes denote ImageReward, PickScore, HPSv2, and Aesthetic score computed on DrawBench generations. "$*$" marks results reproduced by us. "+ SRR" denotes inference-time self-reflective refinement. **Bold** indicates the best among UMMs.

Model	TIIF-Bench ↑		WISE ↑	DPGBench ↑	GenEval ↑	Human Preference* ↑			
	S	L				IR	PS	HPS	Aes
Generation-Only Models									
SD3 Medium (Esser et al., 2024)	64.79	64.79	0.39	84.08	74.00	0.94	22.47	0.29	5.41
FLUX.1-dev (Black Forest Labs, 2024)	66.24	66.72	0.42	84.00	82.00	1.03	22.88	0.30	5.78
Unified Multimodal Models									
Show-o (Xie et al., 2025)	57.34	59.24	0.35	67.27	68.00	0.85	22.00	0.27	5.62
Janus-Pro-7B (Chen et al., 2025b)	65.38	61.10	0.33	84.19	80.00	0.83	22.09	0.27	5.47
BAGEL* (Deng et al., 2025)	74.66	75.42	0.38	84.93	76.77	1.02	22.72	0.29	5.66
AlphaGRPO* (Huang et al., 2026a)	76.96	78.19	0.39	86.05	77.17	0.96	22.63	0.29	5.54
+ SRR	77.60	78.06	–	85.90	79.59	1.18	22.63	0.30	5.54
Ours*	**83.35**	**83.28**	**0.41**	**88.77**	**85.04**	**1.25**	**22.90**	**0.31**	**5.68**



Figure 4: **Qualitative comparisons.** The examples cover complex composition, attribute binding, spatial relations, object counting, aesthetic quality, and prompt adherence. +SRR denotes AlphaGRPO with inference-time self-reflective refinement. T2I-OPSD requires only a single round of prompt-only generation, without inference-time reflection or image regeneration.

mixing coefficient to $\beta = 0.7$. Starting from a pool of visual primitives, including objects, attributes, and spatial relations, we define compositional tasks following the TIIF-Bench taxonomy (Wei et al., 2025) and use Qwen3-235B-A22B (Qwen Team, 2025) to synthesize training prompts by stochastically sampling primitives across three difficulty tiers (Huang et al., 2026a). To avoid overlap with the prompts used during AlphaGRPO training, we independently generate a new training prompt set using the same procedure. Training rollouts use a resolution of 512×512, 40 sampling steps, and a timestep shift of 3.0. We set the text CFG scale to 4.0 and the image CFG scales to 1.0 and 2.0 for policy rollouts and the reflection-conditioned self-teacher branch, respectively. All experiments are run on A100 GPUs with a learning rate of 1×10^{-4}, batch size of 32, and 200 optimization steps.

Benchmarks. We evaluate on four text-to-image benchmarks. GenEval (Ghosh et al., 2023) measures compositional generation; DPG-Bench (Hu et al., 2024) measures semantic alignment under dense, long prompts; TIIF-Bench (Wei et al., 2025) measures instruction following under both short- and long-prompt settings; and WISE (Niu et al., 2025) measures world-knowledge reasoning, with all WISE results evaluated under its updated protocol. Beyond benchmark scores, we generate 200

images on DrawBench (Saharia et al., 2022) for each model and compute ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), and the aesthetic score under a unified evaluation setup to assess the trade-off between compositionality and image quality. For the BAGEL family, we use a resolution of 1024×1024, 50 sampling steps, and a CFG scale of 4.0. Other baselines use their official checkpoints and recommended inference configurations. T2I-OPSD performs only single-round prompt-only generation, without inference-time multi-round refinement, whereas AlphaGRPO+SRR serves as a reference method with additional inference-time refinement.

Baselines. We compare against two groups of baselines. The first group consists of generation-only models, including SD3 (Esser et al., 2024) and FLUX (Black Forest Labs, 2024). The second group consists of unified multimodal models, including Show-o (Xie et al., 2025), Janus-Pro (Chen et al., 2025b), BAGEL (Deng et al., 2025), AlphaGRPO, and AlphaGRPO+SRR (Huang et al., 2026a). Among them, BAGEL, AlphaGRPO, AlphaGRPO+SRR, and T2I-OPSD share the same backbone and form our primary controlled comparison. External baselines use their official checkpoints and officially recommended inference protocols. Unless marked as reproduced by us, benchmark scores are taken from the corresponding papers and published benchmark comparisons. Detailed benchmark comparisons are provided in App. H.

4.2 Quantitative Results

As shown in Table 1, T2I-OPSD delivers broad improvements on compositional generation benchmarks, achieving $83.35/83.28$ on TIIF-Bench-S/L, 88.77 on DPG-Bench, and 85.04 on GenEval, all of which are the best results among the compared methods. Relative to the same-backbone AlphaGRPO model, these scores represent gains of $6.39/5.09$, 2.72, and 7.87 points, respectively. Moreover, T2I-OPSD remains consistently ahead of AlphaGRPO+SRR despite requiring no inference-time refinement. Under the unified DrawBench evaluation, T2I-OPSD further achieves the highest ImageReward, PickScore, and HPSv2 scores while maintaining an Aesthetic score comparable to the original BAGEL model. Overall, T2I-OPSD substantially strengthens compositional alignment in single-round generation without compromising model-based preference quality.

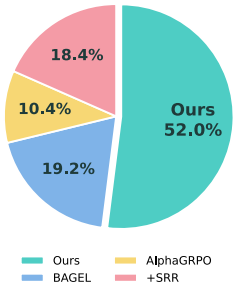


Figure 5: User Study.

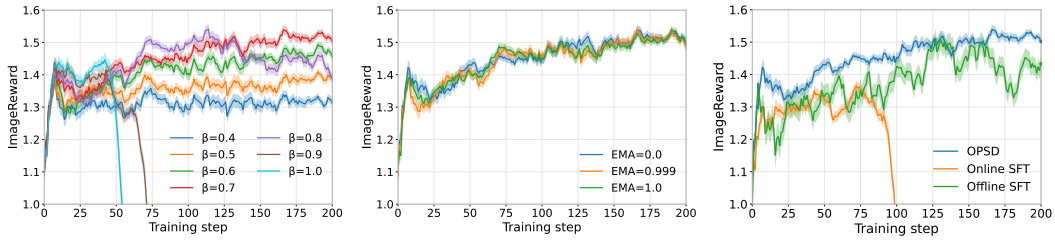
User study. We further evaluate human preference by comparing BAGEL, AlphaGRPO, AlphaGRPO+SRR, and T2I-OPSD. We recruit 20 participants, each of whom evaluates 50 randomly selected image sets and selects their preferred image among the four methods, yielding 1,000 judgments. As shown in Fig. 5, T2I-OPSD receives 52.0% of the votes, compared with 19.2% for BAGEL, 10.4% for AlphaGRPO, and 18.4% for AlphaGRPO+SRR. These results demonstrate a clear human preference for our method, despite requiring no inference-time refinement.

4.3 Qualitative Results

Fig. 4 presents qualitative comparisons of T2I-OPSD with BAGEL, AlphaGRPO, and AlphaGRPO+SRR. In the complex composition example, T2I-OPSD jointly satisfies color and spatial constraints by placing a mint-green deer to the right of a pale-lavender wooden swing and a peach-pink rabbit to its left. T2I-OPSD clearly renders the attribute combination of a purple dog and a black dining table and places the donut to the right of the bench, a spatial relation that AlphaGRPO also achieves after SRR. In the counting example, all three baselines produce extra hangers, whereas T2I-OPSD satisfies the requested count of five. For richer natural scenes, our method also renders fine mask ornamentation and detailed metal and wood textures. These examples complement the quantitative results by illustrating compositional alignment alongside visually detailed image synthesis. Unlike AlphaGRPO+SRR, T2I-OPSD obtains these results in a single round of prompt-only generation, without textual reflection or image regeneration at inference.

4.4 Ablation Studies and Analysis

We ablate the residual weight β and the teacher update strategy, and compare against SFT on teacher-refined images (Fig. 6). We further examine the timestep weighting induced by self-normalized



(a) Residual weight β. (b) Teacher update strategy. (c) Comparison with SFT.

Figure 6: Training dynamics of T2I-OPSD under different residual weights and teacher update strategies, and comparison with online and offline SFT.

endpoint regression in App. E.1. Throughout these experiments, we monitor the student's first-turn generation quality using ImageReward and assess the teacher's relative corrective advantage through teacher win rates (App. E.2), which generally decline over training. These scores are used only for monitoring and do not enter the training objective or sample selection.

Residual weight. As shown in Fig. 6 (a), increasing β does not improve performance monotonically: smaller weights limit the gains from privileged supervision, whereas larger weights yield early improvements but are followed by more pronounced late-stage degradation. We adopt $\beta = 0.7$ as the default setting based on its better overall performance. App. E.2 further shows that $\beta = 0.9$ and 1.0 exhibit faster early declines in teacher win rate, whereas $\beta = 0.7$ slows this decline and retains higher win rates. By limiting the target's deviation from the frozen reference, moderate anchoring reduces the influence of unreliable teacher corrections while retaining useful corrective guidance. This prolongs the stable optimization phase and sustains a higher performance plateau, allowing the student to accumulate improvements rather than deteriorate rapidly as teacher reliability declines.

Teacher update strategy. We compare an online teacher (EMA $= 0$), an EMA teacher (EMA $= 0.999$), and a frozen teacher (EMA $= 1$). In the frozen-teacher setting, both reflection generation and teacher velocity prediction use the frozen initialization, while the current student generates the endpoint images. As shown in Fig. 6 (b), the training curves are nearly identical across the three settings. At $\beta = 0.7$, App. E.2 also shows similar overall declines in teacher win rate across all three update strategies, converging to comparable levels later in training. The decline persists even with a frozen teacher, indicating that the loss of supervision reliability is driven primarily by changes in the distribution of privileged context rather than by teacher parameter updates. This mechanism corresponds to the second term of Eq. 9, where changes in student-generated images refresh the states and privileged context even at fixed teacher parameters.

Comparison with SFT. We compare T2I-OPSD with directly fine-tuning on teacher-refined images, using either a fixed image set (Offline SFT) or images refreshed during training (Online SFT). As shown in Fig. 6 (c), Offline SFT avoids collapse but achieves lower final performance, reflecting the limitations of fixed supervision. Online SFT refreshes its targets but becomes unstable and collapses, showing that target refresh alone is insufficient for sustained improvement. T2I-OPSD instead combines evolving teacher supervision with reference-centered velocity distillation, surpassing the performance plateau of Offline SFT while maintaining more stable training.

5 Conclusion

We presented T2I-OPSD, which distills reflection-conditioned supervision into prompt-only generation through velocity matching at shared student-induced states. We identified co-drift as the coupled degradation of the student and its supervision under endogenous privilege, and introduced reference-centered residual distillation to stabilize supervision. On AlphaGRPO-initialized BAGEL-7B, T2I-OPSD improves compositional alignment while preserving the original single-round inference procedure. These results demonstrate that self-reflection can improve generation through training-time supervision without additional inference-time correction.

6 AI USE STATEMENT

We used generative AI tools to generate synthetic training prompts as described in Section 4.1 and to polish the manuscript's language. We take full responsibility for the final content of this work, including all AI-assisted text and artifacts.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024. Official inference repository for FLUX.1 models.

Black Forest Labs. FLUX.2: Official inference repository, 2025. URL https://github.com/black-forest-labs/flux2.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv:2509.23951*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv:2505.09568*, 2025a.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024a.

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. In *ICLR*, 2024b.

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:2501.17811*, 2025b.

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

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv:2605.12500*, 2026.

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

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

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *NeurIPS*, 2023.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv:2605.08063*, 2026.

Siming Fu, Zheming Fu, Ruizhe He, Hualiang Wang, Jie Huang, Xiaoxiao Ma, Mingchen Zhong, Weihu Huang, Xiaoxuan He, and Haojun Xu. Any-opd: Heterogeneous on-policy distillation for flow-matching models via representation-space bridging. *arXiv:2608.03316*, 2026.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv:2404.14396*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. Suder: Self-improving unified large multimodal models for understanding and generation with dual self-rewards. In *KDD*, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv:2403.05135*, 2024.

Runhui Huang, Jie Wu, Rui Yang, Zhe Liu, and Hengshuang Zhao. Alphagrpo: Unlocking self-reflective multimodal generation in unified multimodal models via decompositional verifiable reward. In *ICML*, 2026a.

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. In *ICLR*, 2026b.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, et al. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv:2605.05204*, 2026.

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. *NeurIPS*, 2025.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. *arXiv:2510.12784*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *NeurIPS*, 2023.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv:2605.15055*, 2026a.

Yuhan Li, Fangao Zeng, Sicong Kang, Mengfei Xu, Hao Zhou, Wei Li, Pipei Huang, and Bingbing Ni. Rl-native distillation: Exploiting scored trajectories for few-step image generation. *arXiv:2608.09226*, 2026b.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv:2405.08748*, 2024.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv:2506.03147*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *NeurIPS*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv:2209.03003*, 2022.

Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning. *arXiv:2505.23380*, 2025.

Shentong Mo and Sukmin Yun. Lvrpo: Language-visual alignment with grpo for multimodal understanding and generation. *arXiv:2603.27693*, 2026.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv:2503.07265*, 2025.

Jiadong Pan, Liang Li, Yuxin Peng, Yu-Ming Tang, Shuohuan Wang, Yu Sun, Hua Wu, Qingming Huang, and Haifeng Wang. Learning to generate via understanding: Understanding-driven intrinsic rewarding for unified multimodal models. *arXiv:2603.06043*, 2026.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv:2504.06256*, 2025.

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. In *ICLR*, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, 2025.

Qwen Team. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

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. *NeurIPS*, 2022.

Stability AI. Introducing stable diffusion 3.5, 2024. URL https://stability.ai/news/introducing-stable-diffusion-3-5.

Jiayang Sun, Pin Wang, Hongbo Wang, Xinyue Liu, Huaibo Huang, and Ran He. Towards fine-grained attribution: Instance-aware preference optimization for aligning diffusion models. In *CVPR*, 2026.

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:2406.06525*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv:2405.09818*, 2024.

Ritesh Thawkar, Shravan Venkatraman, Omkar Thawakar, Abdelrahman Shaker, Fahad Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. Ask, solve, generate: Self-evolving unified multimodal understanding and generation via self-consistency rewards. *arXiv:2606.27376*, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In *ICCV*, 2025.

Dianyi Wang, Ruihang Li, Feng Han, Chaofan Ma, Wei Song, Siyuan Wang, Yibin Wang, Yi Xin, Hongjian Liu, Zhixiong Zhang, et al. Deepgen 1.0: A lightweight unified multimodal model for advancing image generation and editing. *arXiv:2602.12205*, 2026.

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:2409.18869*, 2024.

Qingyan Wei, Guangzhao Li, Xiaobing Tu, Yinggui Wang, Xiantao Zhang, Jinkui Ren, Xiaohong Liu, and Linfeng Zhang. Step-opd: Rethinking output targets and internal dynamics in on-policy distillation for diffusion models. *arXiv:2608.04887*, 2026.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv:2508.02324*, 2025a.

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 *CVPR*, 2025b.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *CVPR*, 2026a.

Junlong Wu, Jiuzhou Lin, Jia Sun, Boheng Zhang, Huaiqing Wang, Dewen Fan, Houde Liu, Qianqian Gan, Fan Yang, and Tingting Gao. Step back to move forward: Reflection-aware preference optimization for visual generation. *arXiv:2609.04282*, 2026b.

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:2306.09341*, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv:2505.07818*, 2025.

Sen Ye, Mengde Xu, Shuyang Gu, Di He, Liwei Wang, and Winston Hu. Understanding vs. generation: Navigating optimization dilemma in multimodal models. In *ICLR*, 2026.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv:2511.22699*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv:2601.18734*, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *ICLR*, 2026.

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. In *ICLR*, 2025.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, et al. On-policy self-distillation in diffusion models. *arXiv:2608.24646*, 2026.

A Implementation Details

We initialize BAGEL-7B-MoT from the reflection-capable AlphaGRPO checkpoint and train it with LoRA. The student and self-teacher share the current parameters, while the reference branch retains the frozen initial SRR adapter. We train on eight A100 GPUs with one example per GPU and four gradient accumulation steps, yielding a global batch size of 32. For each generated image, we perform a 40-step rollout and use 24 timesteps for distillation. At inference, we retain only the prompt-only student, without reflection or a second round of image generation. Table 2 summarizes the full training and inference configurations.

Table 2: Training and inference configurations of T2I-OPSD.

Setting	T2I-OPSD
Backbone	BAGEL-7B-MoT
Initialization	AlphaGRPO checkpoint
Hardware	$8\times$ A100 GPUs
LoRA rank	32
LoRA alpha	64
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	1×10^{-4}
Learning rate schedule	Constant
Weight decay	1×10^{-4}
Gradient clipping	1.0 (global norm)
Training precision	BF16 (bfloat16 autocast)
Training steps	200
Global batch size	32
Training timesteps per image	24
Residual weight β	0.7
Teacher update strategy	Online (shared parameters, EMA decay $= 0$)
Reference update strategy	Frozen (SRR adapter)
Training resolution	512×512
Rollout sampling steps	40
Rollout timestep shift	3.0
Text CFG scale	4.0
Image CFG scale (rollout)	1.0
Image CFG scale (self-teacher)	2.0
Inference resolution	1024×1024
Inference sampling steps	50
Inference CFG scale	4.0
Inference timestep shift	3.0

B Limitations

T2I-OPSD depends strongly on the reflection capability of its initialization. The model must already be able to identify generation errors and provide useful corrections to serve as its own teacher. For backbones without these capabilities, reflection must be learned before distillation. In this work, we meet this prerequisite by initializing from an AlphaGRPO checkpoint. This dependence extends beyond the quality of reflection to the preferences it encodes. Errors that the model recognizes and prioritizes receive corrective supervision, while those it overlooks receive little useful guidance. The emphasis and blind spots of reflection therefore shape the capabilities that on-policy self-distillation improves. Our substantial gains in compositional alignment, alongside more limited gains on WISE, are consistent with an initialization whose reflection focuses primarily on prompt alignment rather than world knowledge.

C Training Algorithm

The algorithm below summarizes the T2I-OPSD training procedure. For each prompt, the current model generates an image and its reflection once, then reuses them as privileged context across K sampled timesteps. At each timestep, we add fresh noise to the generated endpoint and evaluate the student, self-teacher, and frozen reference at the resulting shared state. The student and self-teacher share the current parameters θ, while the reference retains θ_0. We interpolate the teacher and reference velocities, convert the result into an endpoint target, and optimize the prompt-only student using the self-normalized objective in Eq. (13). During optimization, we keep the sampled states fixed and stop gradients through the supervision targets.

Algorithm 1 Text-to-Image On-Policy Self-Distillation (T2I-OPSD)

Require: Reflection-capable initialization θ_0, prompt dataset $\mathcal{D}$, timestep distribution $p(t)$, fixed mixing coefficient β, learning rate η, and K timesteps per image.
Initialize: Student $\theta \leftarrow \theta_0$; keep reference θ_0 frozen.
for each training iteration **do**
  Sample a minibatch $\mathcal{B}$ from $\mathcal{D}$ and set $\mathcal{L} \leftarrow 0$.
  for each prompt $c \in \mathcal{B}$ **do**
    Sample $\xi \sim \mathcal{N}(0, I)$ and generate $\hat{x}_0 \leftarrow \mathrm{sg}[\mathrm{Sample}_\theta(\xi; c)]$.
    Generate $r \leftarrow \mathcal{R}_\theta(c, \hat{x}_0)$ without gradients, with $c_S \leftarrow c$ and $c_T \leftarrow (c, \hat{x}_0, r)$.
    Sample K timesteps $t_1, \ldots, t_K$ according to $p(t)$.
    for each sampled timestep t **do**
      Sample fresh $\epsilon \sim \mathcal{N}(0, I)$, independently of ξ, and set $x_t \leftarrow \mathrm{sg}[(1-t)\hat{x}_0 + t\epsilon]$.
      Compute $v_{\mathrm{stu}} \leftarrow v_\theta(x_t, t \mid c_S)$.
      Compute $v_{\mathrm{tea}} \leftarrow \mathrm{sg}[v_\theta(x_t, t \mid c_T)]$ and $v_{\mathrm{anc}} \leftarrow \mathrm{sg}[v_{\theta_0}(x_t, t \mid c_S)]$.
      Form $v_{\mathrm{target}} \leftarrow v_{\mathrm{anc}} + \beta(v_{\mathrm{tea}} - v_{\mathrm{anc}})$.
      Convert to $x_\theta \leftarrow x_t - t v_{\mathrm{stu}}$ and $x_{\mathrm{target}} \leftarrow \mathrm{sg}[x_t - t v_{\mathrm{target}}]$.
      Accumulate $\mathcal{L} \leftarrow \mathcal{L} + \dfrac{1}{|\mathcal{B}|K} \dfrac{\|x_\theta - x_{\mathrm{target}}\|_2^2}{\mathrm{sg}[\mathrm{mean}(|x_\theta - x_{\mathrm{target}}|)]}$.
    end for
  end for
  Update the trainable student parameters $\theta \leftarrow \mathrm{AdamW}(\theta, \nabla_\theta \mathcal{L}; \eta)$.
end for
Output: Prompt-only student $v_\theta(x_t, t \mid c)$.

D Motivation for Forward Resampling

AlphaGRPO trains reflective refinement with independently sampled noise. We observe that reusing the initial noise from the first generation can introduce pronounced visual artifacts during reflective refinement. As shown in Fig. 7, noise reuse produces unnatural high-frequency textures on the wall, bench, and ground, whereas fresh noise yields more natural surface textures and a different scene layout. This observation indicates that its reflective refinement capability is sensitive to noise reuse and motivates our use of independent noise to construct distillation states. Specifically, we forward-noise the current student's generated endpoint to obtain $x_t = \alpha_t \hat{x}_0 + \sigma_t \epsilon$, where ϵ is independent of the initial generation noise ξ, rather than tying the distillation states to the original generation trajectory.

E Additional Ablation Studies

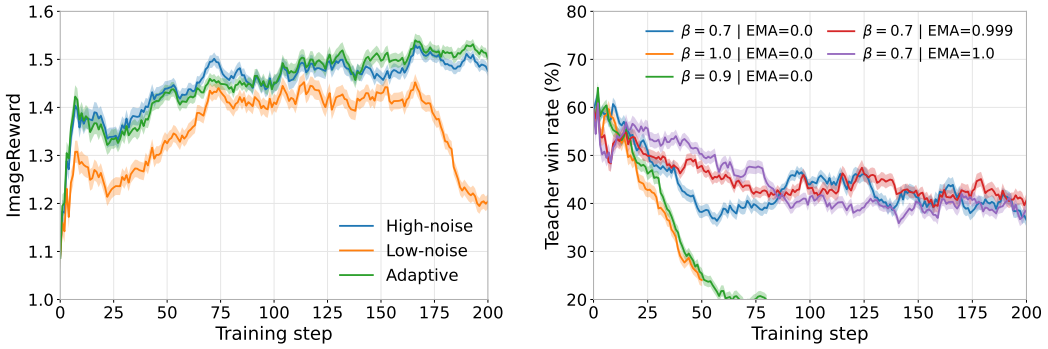
E.1 Timestep Credit Assignment

We compare three strategies for timestep credit assignment: High-noise, Low-noise, and our Adaptive strategy. High-noise uses the noise variance σ_t^2 as a fixed weight, emphasizing high-noise states early in denoising. Low-noise uses $1 - \sigma_t^2$, emphasizing low-noise states later in denoising. Here, early and late refer to the denoising stages within a generation rather than training iterations. Unlike these preset weights, Adaptive uses the self-normalized endpoint regression objective in the main text to adjust the supervision strength according to prediction errors. As shown in Fig. 8 (a), Low-noise yields lower ImageReward scores throughout most of training and degrades sharply at



(a) First-turn image. (b) SRR with reused noise. (c) SRR with fresh noise.

Figure 7: **Effect of noise reuse on reflective refinement in AlphaGRPO.** (a) The first-turn image. (b) SRR with the initial noise reused from the first generation. (c) SRR with independently sampled fresh noise. Reusing the initial noise introduces pronounced texture artifacts on the wall, bench, and ground, whereas fresh noise produces more natural surface textures.



(a) Timestep credit assignment. (b) Teacher win rate.

Figure 8: Training dynamics of timestep weighting strategies and the teacher's relative corrective advantage. (a) ImageReward curves for High-noise, Low-noise, and Adaptive. (b) Teacher win rates under different residual weights and teacher update strategies, with faint traces showing raw measurements and solid curves showing smoothed trends.

later training steps. High-noise and Adaptive follow similar overall trends, but Adaptive generally maintains slightly higher scores toward the end of training. These results support adaptive timestep weighting rather than a fixed emphasis on a particular denoising stage.

E.2 Teacher Win Rate Analysis

Evaluation protocol. We use teacher win rate to measure the corrective advantage of reflection conditioning over prompt-only conditioning. After obtaining the current student's first-turn image and its reflection, we independently sample fresh noise and use the same noise for both the prompt-only student branch and the teacher branch conditioned on the image and reflection. Teacher win rate is the fraction of pairs in which the teacher image receives a strictly higher ImageReward score than the student image. This metric is used only for monitoring and analysis and does not affect training.

Training dynamics. As shown in Fig. 8 (b), teacher win rates generally decline over training, indicating that reflective corrections become less advantageous relative to the student. With $\beta = 0.9$ or 1.0, win rates fall more rapidly at early training steps, consistent with the training degradation under weak anchoring reported in the main text. In contrast, $\beta = 0.7$ slows this decline and retains higher

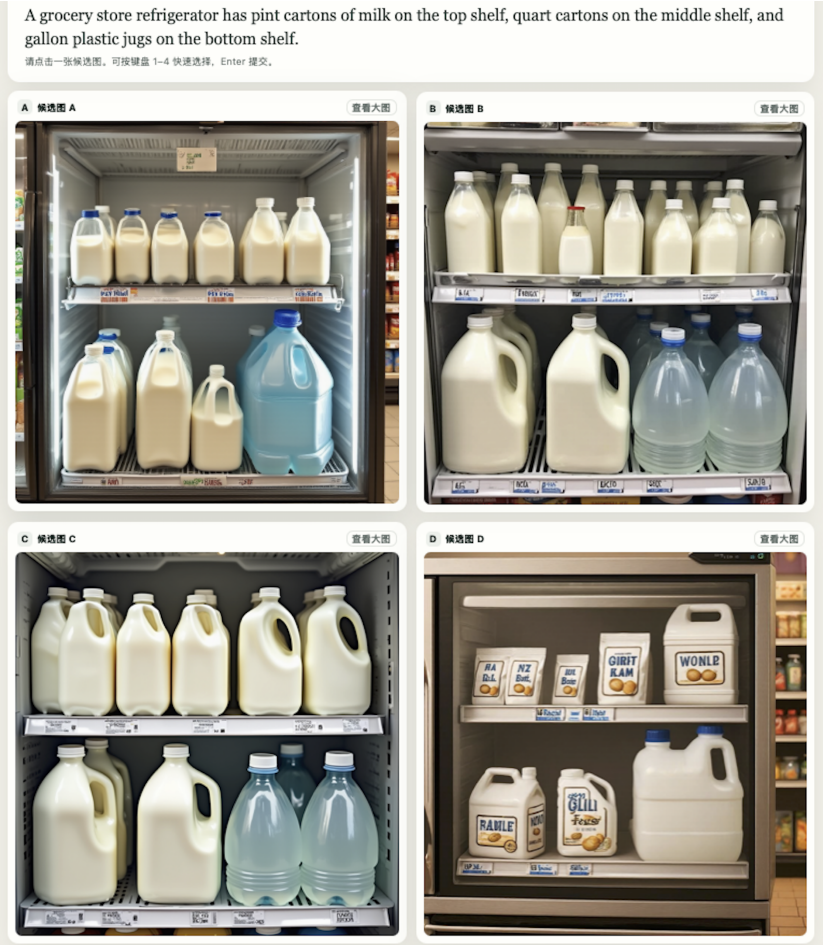


Figure 9: **User study interface.** The prompt appears above four anonymous images arranged in random order. Participants select their preferred image based primarily on prompt alignment, while also considering aesthetic quality.

teacher win rates during subsequent training, supporting the role of moderate reference anchoring in prolonging effective supervision.

At $\beta = 0.7$, online, EMA, and frozen teachers all exhibit declining win rates and reach comparable levels later in training. The decline persists even when teacher parameters are fixed, showing that freezing parameters alone is insufficient to preserve the teacher's relative corrective advantage. Student-generated images continue to refresh the supervised states and privileged context, so fixed parameters do not eliminate the cross-step feedback between the student and its supervision. This is consistent with our analysis of endogenous privilege and co-drift in the main text.

F User Study

Fig. 9 shows the interface used in our user study. For each prompt, participants view four images generated by BAGEL, AlphaGRPO, AlphaGRPO+SRR, and T2I-OPSD. We display the prompt above the images and randomly assign the four methods to positions A–D without revealing their identities. Participants select the image they prefer, prioritizing alignment with the prompt while also considering aesthetic quality. They can enlarge each image to inspect its details before making a choice.

G MORE VISUALIZATION

Additional qualitative analysis. Fig. 10 compares BAGEL, AlphaGRPO, AlphaGRPO with SRR, and T2I-OPSD on compositional prompts. In the first row, our method preserves both the brown bed and the pink cell phone; in the second row, it places the knife to the right of the suitcase, matching the requested spatial relation. The last row further illustrates its ability to combine metallic toiletries with a pink grapefruit half on a patterned plate. These examples complement the quantitative results by illustrating attribute binding and spatial alignment in single-round generation, without inference-time refinement.

H ADDITIONAL BENCHMARK RESULTS

Additional quantitative analysis. The detailed results highlight the compositional gains of T2I-OPSD. On GenEval (Table 3), our method achieves the highest overall score among the compared UMMs (85.04) and a color-attribution score of 82.00, exceeding AlphaGRPO with SRR by 17.00 points. On DPG-Bench (Table 8), it attains the best entity and attribute scores among the compared UMMs (92.75 and 93.16), with an overall score of 88.77. On TIIF-Bench (Tables 4–7), it leads the compared UMMs in text generation (78.28/77.38 for short/long prompts) and real-world following (89.93/87.31), without inference-time refinement. On WISE (Table 9), its overall score improves from 0.38/0.39 for BAGEL/AlphaGRPO to 0.41. Our method primarily targets text–image alignment and does not introduce additional world knowledge. The modest WISE improvement is consistent with this focus, while a gap to stronger models on knowledge-intensive generation remains.

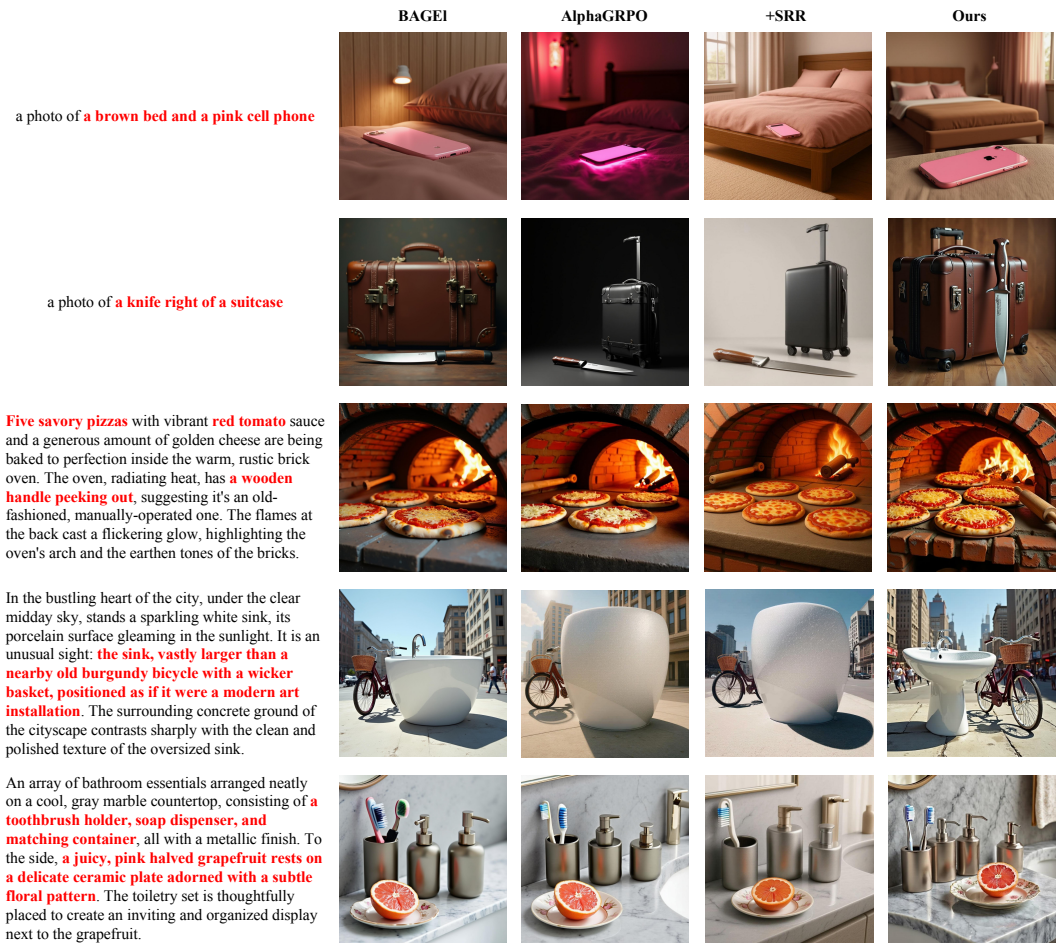


Figure 10: Additional qualitative comparisons for text-to-image generation. Each row shows a prompt and the results of BAGEL, AlphaGRPO, AlphaGRPO with SRR, and our method, from left to right. Red text highlights selected requirements in the prompts.

Table 3: GenEval results (%). † denotes an LLM prompt rewriter. All metrics are higher-is-better; – indicates unavailable results. **Bold** indicates the best among UMMs. "∗" marks results reproduced by us, including the + SRR variant under AlphaGRPO. "+ SRR" denotes inference-time self-reflective refinement.

Model	Single	Two	Count	Colors	Position	Color attr.	Overall
Generation-Only Models							
PixArt-α (Chen et al., 2024b)	98.00	50.00	44.00	80.00	8.00	7.00	48.00
Emu3-Gen (Wang et al., 2024)	98.00	71.00	34.00	81.00	17.00	21.00	54.00
SDXL (Podell et al., 2024)	98.00	74.00	39.00	85.00	15.00	23.00	55.00
DALL-E 3 (Betker et al., 2023)	96.00	87.00	47.00	83.00	43.00	45.00	67.00
SD3-Medium (Esser et al., 2024)	99.00	94.00	72.00	89.00	33.00	60.00	74.00
FLUX.1-dev† (Black Forest Labs, 2024)	98.00	93.00	75.00	93.00	68.00	65.00	82.00
Unified Multimodal Models							
SEED-X (Ge et al., 2024)	97.00	58.00	26.00	80.00	19.00	14.00	49.00
TokenFlow-XL (Qu et al., 2025)	95.00	60.00	41.00	81.00	16.00	24.00	55.00
ILLUME (Wang et al., 2025)	99.00	86.00	45.00	71.00	39.00	28.00	61.00
Transfusion (Zhou et al., 2025)	–	–	–	–	–	–	63.00
Emu3-Gen† (Wang et al., 2024)	99.00	81.00	42.00	80.00	49.00	45.00	66.00
Show-o (Xie et al., 2025)	98.00	80.00	66.00	84.00	31.00	50.00	68.00
Janus-Pro-7B (Chen et al., 2025b)	99.00	89.00	59.00	90.00	**79.00**	66.00	80.00
MetaQuery-XL† (Pan et al., 2025)	–	–	–	–	–	–	80.00
BAGEL∗ (Deng et al., 2025)	**100.00**	**97.98**	73.75	81.91	50.00	57.00	76.77
AlphaGRPO∗ (Huang et al., 2026a)	**100.00**	89.90	**80.00**	85.11	54.00	54.00	77.17
+ SRR (Huang et al., 2026a)	98.75	92.93	72.50	89.36	59.00	65.00	79.59
Ours∗	**100.00**	**97.98**	78.75	**91.49**	60.00	**82.00**	**85.04**

Table 4: TIIF-Bench with short prompts: overall, basic following, and real-world results (%). All metrics are higher-is-better. **Bold** indicates the best among UMMs. "∗" marks results reproduced by us, including the + SRR variant under AlphaGRPO. "+ SRR" denotes inference-time self-reflective refinement.

Model	Overall	Basic avg.	Attribute	Relation	Reasoning	Real world
Generation-Only Models						
FLUX.1 dev (Black Forest Labs, 2024)	71.09	83.12	87.05	87.25	75.01	70.72
SD XL (Podell et al., 2024)	54.96	65.72	59.33	77.57	60.32	50.92
SD 3 (Esser et al., 2024)	67.46	78.32	83.33	82.07	71.07	63.23
SD 3.5 L (Stability AI, 2024)	71.15	78.34	79.50	80.96	72.46	64.43
Llamagen (Sun et al., 2024)	41.67	53.00	48.33	59.57	51.07	39.73
Infinity (Han et al., 2025)	62.07	73.08	74.33	72.82	72.07	54.28
Unified Multimodal Models						
Show-o (Xie et al., 2025)	59.72	73.08	74.83	78.82	65.57	55.02
JanusPro (Chen et al., 2025b)	66.50	79.33	79.33	78.32	80.32	65.84
BAGEL∗ (Deng et al., 2025)	74.66	83.49	83.50	88.70	78.25	76.49
AlphaGRPO∗ (Huang et al., 2026a)	76.96	83.77	83.00	87.02	81.29	78.73
+ SRR (Huang et al., 2026a)	77.60	**88.46**	89.00	87.26	**89.10**	76.12
Ours∗	**83.35**	87.83	**92.50**	**91.13**	79.85	**89.93**

Table 5: TIIF-Bench with short prompts: advanced following results (%). All metrics are higher-is-better. **Bold** indicates the best among UMMs. Attr., Rel., and Reas. denote attribute, relation, and reasoning. "$*$" marks results reproduced by us, including the + SRR variant under AlphaGRPO. "+ SRR" denotes inference-time self-reflective refinement.

Model	Adv. avg.	Attr.+Rel.	Attr.+Reas.	Rel.+Reas.	Style	Text
Generation-Only Models						
FLUX.1 dev (Black Forest Labs, 2024)	65.79	67.07	73.84	69.09	66.67	43.83
SD XL (Podell et al., 2024)	49.73	47.82	56.22	52.59	73.33	16.83
SD 3 (Esser et al., 2024)	61.46	61.07	68.84	50.96	66.67	59.83
SD 3.5 L (Stability AI, 2024)	67.67	66.46	73.53	60.03	73.33	70.52
Llamagen (Sun et al., 2024)	35.89	38.82	40.84	49.59	46.67	0.00
Infinity (Han et al., 2025)	56.64	60.44	74.22	60.22	80.00	10.83
Unified Multimodal Models						
Show-o (Xie et al., 2025)	53.67	60.95	68.59	66.46	63.33	3.83
JanusPro (Chen et al., 2025b)	59.71	66.07	70.46	67.22	60.00	28.83
BAGEL* (Deng et al., 2025)	72.51	77.11	67.23	75.72	73.33	51.58
AlphaGRPO* (Huang et al., 2026a)	75.93	83.20	72.13	74.91	76.67	55.66
+ SRR (Huang et al., 2026a)	**81.19**	85.55	**83.81**	**81.24**	**90.00**	16.29
Ours*	79.46	**85.63**	79.02	73.83	80.00	**78.28**

Table 6: TIIF-Bench with long prompts: overall, basic following, and real-world results (%). All metrics are higher-is-better. **Bold** indicates the best among UMMs. "$*$" marks results reproduced by us, including the + SRR variant under AlphaGRPO. "+ SRR" denotes inference-time self-reflective refinement.

Model	Overall	Basic avg.	Attribute	Relation	Reasoning	Real world
Generation-Only Models						
FLUX.1 dev (Black Forest Labs, 2024)	71.78	78.65	83.17	80.39	72.39	71.47
SD XL (Podell et al., 2024)	42.13	53.28	50.83	62.57	46.57	41.59
SD 3 (Esser et al., 2024)	66.09	77.75	79.83	78.82	74.07	67.34
SD 3.5 L (Stability AI, 2024)	66.96	79.56	76.50	83.21	78.71	66.39
Llamagen (Sun et al., 2024)	38.22	50.00	42.33	60.32	47.32	35.62
Infinity (Han et al., 2025)	62.32	75.41	76.83	77.57	71.82	56.89
Unified Multimodal Models						
Show-o (Xie et al., 2025)	58.86	75.83	79.83	78.32	69.32	50.92
JanusPro (Chen et al., 2025b)	65.02	78.25	82.33	73.32	79.07	60.25
BAGEL* (Deng et al., 2025)	75.42	81.40	82.50	83.44	78.25	77.24
AlphaGRPO* (Huang et al., 2026a)	78.19	83.55	88.00	86.91	75.73	80.60
+ SRR (Huang et al., 2026a)	78.06	**86.47**	87.00	85.92	**86.50**	71.64
Ours*	**83.28**	85.96	**88.50**	**88.01**	81.38	**87.31**

Table 7: TIIF-Bench with long prompts: advanced following results (%). All metrics are higher-is-better. **Bold** indicates the best among UMMs. Attr., Rel., and Reas. denote attribute, relation, and reasoning. "$*$" marks results reproduced by us, including the + SRR variant under AlphaGRPO. "+ SRR" denotes inference-time self-reflective refinement.

Model	Adv. avg.	Attr.+Rel.	Attr.+Reas.	Rel.+Reas.	Style	Text
Generation-Only Models						
FLUX.1 dev (Black Forest Labs, 2024)	68.54	73.69	73.34	71.59	66.67	52.83
SD XL (Podell et al., 2024)	36.22	35.57	45.34	36.09	60.00	0.83
SD 3 (Esser et al., 2024)	59.56	64.07	70.34	57.84	76.67	20.83
SD 3.5 L (Stability AI, 2024)	61.18	61.89	74.15	61.53	63.33	42.52
Llamagen (Sun et al., 2024)	32.61	31.57	47.22	46.22	33.33	0.00
Infinity (Han et al., 2025)	54.98	55.57	64.71	59.71	73.33	23.83
Unified Multimodal Models						
Show-o (Xie et al., 2025)	50.38	56.82	68.96	56.22	66.67	2.83
JanusPro (Chen et al., 2025b)	58.82	56.20	70.84	59.97	70.00	33.83
BAGEL* (Deng et al., 2025)	75.31	81.06	76.05	71.56	66.67	61.99
AlphaGRPO* (Huang et al., 2026a)	79.23	82.45	78.70	79.91	76.67	54.75
+ SRR (Huang et al., 2026a)	79.81	82.26	80.22	80.89	**83.33**	44.80
Ours*	**82.94**	**84.16**	**83.60**	**82.53**	76.67	**77.38**

Table 8: DPG-Bench category-level results (%). All metrics are higher-is-better; – indicates unavailable results. **Bold** indicates the best among UMMs. '$*$ marks results reproduced by us, including the + SRR variant under AlphaGRPO. '+ SRR denotes inference-time self-reflective refinement.

Model	Global	Entity	Attribute	Relation	Other	Overall
Generation-Only Models						
Hunyuan-DiT (Li et al., 2024)	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ (Chen et al., 2024a)	86.89	82.89	88.94	86.59	87.68	80.54
DALLE3 (Betker et al., 2023)	90.97	89.61	88.39	90.58	89.83	83.50
SD3-medium (Esser et al., 2024)	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev (Black Forest Labs, 2024)	82.10	89.50	88.70	91.10	89.40	84.00
OmniGen (Xiao et al., 2025)	87.90	88.97	88.47	87.95	83.56	81.16
Unified Multimodal Models						
Show-o (Xie et al., 2025)	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 (Wang et al., 2024)	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL (Qu et al., 2025)	78.72	79.22	81.29	85.22	71.20	73.38
Janus (Wu et al., 2025b)	82.33	87.38	87.70	85.46	86.41	79.68
Janus Pro (Chen et al., 2025b)	86.90	88.90	89.40	89.32	89.48	84.19
BLIP3-o 4B (Chen et al., 2025a)	–	–	–	–	–	79.36
BLIP3-o 8B (Chen et al., 2025a)	–	–	–	–	–	81.60
UniWorld-V1 (Lin et al., 2025)	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 (Wu et al., 2026a)	88.81	88.83	90.18	89.37	90.27	83.57
BAGEL* (Deng et al., 2025)	87.85	90.58	90.71	89.95	90.54	84.93
AlphaGRPO* (Huang et al., 2026a)	**91.81**	90.79	90.89	90.91	90.48	86.05
+ SRR (Huang et al., 2026a)	88.64	91.27	91.26	88.13	89.90	85.90
Ours*	87.35	**92.75**	**93.16**	**91.87**	**90.63**	**88.77**

Table 9: WISE results for open-weight models under the WISE_Verified protocol, on a 0–1 scale. CoT denotes chain-of-thought prompting. All metrics are higher-is-better. **Bold** indicates the best among the listed models. $*$ marks results reproduced by us.

Model	Culture	Time	Space	Biology	Physics	Chem.	Overall
SenseNova-U1-A3B-MoT (w/ CoT) (Diao et al., 2026)	**0.83**	**0.78**	**0.92**	**0.84**	**0.80**	**0.84**	**0.83**
SenseNova-U1-8B-MoT (w/ CoT) (Diao et al., 2026)	0.80	0.74	0.86	0.76	0.78	0.80	0.79
SenseNova-U1-A3B-MoT (Diao et al., 2026)	0.75	0.63	0.74	0.68	0.77	0.68	0.72
SenseNova-U1-8B-MoT (Diao et al., 2026)	0.69	0.60	0.74	0.63	0.72	0.69	0.68
BAGEL (w/ CoT) (Deng et al., 2025)	0.78	0.63	0.57	0.38	0.55	0.51	0.63
DeepGen 1.0 (Wang et al., 2026)	0.65	0.41	0.72	0.39	0.59	0.45	0.57
FLUX.2-dev (Black Forest Labs, 2025)	0.67	0.57	0.66	0.37	0.53	0.38	0.57
Qwen-Image (Wu et al., 2025a)	0.63	0.53	0.56	0.34	0.48	0.25	0.51
Z-Image (Z-Image Team, 2025)	0.55	0.47	0.51	0.33	0.48	0.18	0.45
FLUX.2-klein-9B (Black Forest Labs, 2025)	0.49	0.39	0.55	0.38	0.48	0.23	0.44
HunyuanImage-3.0 (Cao et al., 2025)	0.53	0.39	0.48	0.31	0.45	0.24	0.44
Uniworld_V1 (Lin et al., 2025)	0.52	0.49	0.55	0.23	0.40	0.17	0.43
FLUX.1-dev (Black Forest Labs, 2024)	0.52	0.40	0.53	0.18	0.38	0.24	0.42
SD-3.5-large (Stability AI, 2024)	0.49	0.41	0.44	0.30	0.38	0.21	0.40
FLUX.2-klein-4B (Black Forest Labs, 2025)	0.44	0.37	0.47	0.32	0.39	0.33	0.40
SD-3-medium (Esser et al., 2024)	0.47	0.41	0.40	0.20	0.38	0.26	0.39
SD-3.5-medium (Stability AI, 2024)	0.48	0.38	0.38	0.18	0.39	0.20	0.38
FLUX.1-schnell (Black Forest Labs, 2024)	0.47	0.33	0.47	0.21	0.38	0.10	0.36
SD-XL-0.9 (Podell et al., 2024)	0.49	0.37	0.24	0.27	0.33	0.18	0.36
Janus-Pro-7B (Chen et al., 2025b)	0.37	0.35	0.28	0.28	0.40	0.23	0.33
SD-1.5 (Rombach et al., 2022)	0.45	0.31	0.21	0.21	0.22	0.15	0.31
Janus-1.3B (Wu et al., 2025b)	0.32	0.28	0.18	0.23	0.34	0.18	0.27
Janus-Pro-1B (Chen et al., 2025b)	0.31	0.23	0.23	0.22	0.31	0.20	0.27
BAGEL* (Deng et al., 2025)	0.45	0.41	0.35	0.26	0.42	0.21	0.38
AlphaGRPO* (Huang et al., 2026a)	0.45	0.47	0.39	0.27	0.42	0.22	0.39
Ours*	0.51	0.47	0.32	0.29	0.44	0.17	0.41