DUETGRPO: CO-ADAPTIVE JOINT REINFORCEMENT LEARNING FOR REASONING-DRIVEN IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint training allows the language and visual policies in reasoning-driven image generation to co-adapt prompt formulation and image synthesis. However, existing joint training faces a central challenge: relying solely on image rewards induces reward mismatch, as rendering quality is an unreliable proxy for reasoning and prompt correctness. In this paper, we propose DUETGRPO, a joint reinforcement learning framework that combines modality-specific optimization with shared alignment feedback to support online co-adaptation of language and visual policies. To supervise language outputs directly, we use a text-only evaluator to assess reasoning correctness and check generation prompts against a candidate-independent checklist of instruction requirements. We route this text reward exclusively to the language policy and share image–text alignment feedback across both policies, aligning their joint updates with distinct objectives: the language policy learns to reason correctly and formulate prompts that satisfy the user instruction and can be faithfully rendered, while the visual policy learns to follow those prompts. Experiments improve WISE Overall from 0.72 to 0.77 and T2I-CoReBench Overall from 59.1 to 60.7. Improvements are also observed when this recipe transfers from modular MLLM + diffusion transformer (DiT) pipelines to a mixture-of-transformers (MoT) architecture.

1 INTRODUCTION

Unified multimodal models (UMMs) (Zhao et al., 2025; Liao et al., 2025) support reasoning-driven image generation through a language policy that reasons about a user instruction and produces a generation prompt—a text description of the desired image—and a visual policy that renders it (Jiao et al., 2025; Kou et al., 2026). Figure 1 illustrates this reasoning-to-generation process across physical, temporal, spatial, and chemical requirements. Updating either policy alone limits their mutual adaptation, motivating joint reinforcement learning (RL) that updates both from fresh trajectories in a shared online loop.

Joint RL must preserve reasoning and prompt correctness while allowing the two policies to co-adapt. Existing methods guide language updates mainly through image rewards: PromptRL (Wang et al., 2026b) shares them across both policies, while UniGRPO (Liu et al., 2026) and Dual-GRPO (Kou et al., 2026) reward alignment between generated images and the original user instructions. Such feedback supports coordination, but its dependence on the renderer makes it an indirect measure of reasoning correctness: a low image score may reflect a rendering failure despite correct reasoning, while a high score need not validate the reasoning that preceded it. These limitations motivate checking the reasoning and generation prompt directly against the user instruction. Yet the same instruction can admit multiple valid reasoning paths and generation prompts. Evaluation must therefore assess task requirements while accommodating diverse valid solutions, consistent with PromptEnhancer's fine-grained evaluation (Wang et al., 2026c). Joint RL must combine direct feedback on reasoning and prompt correctness with image-based feedback that helps the language and visual policies adapt to one another.

We systematically investigate how text rewards, image–text alignment rewards, and gradient routing shape joint optimization. We first study how to reward open-ended language outputs by assess-

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.

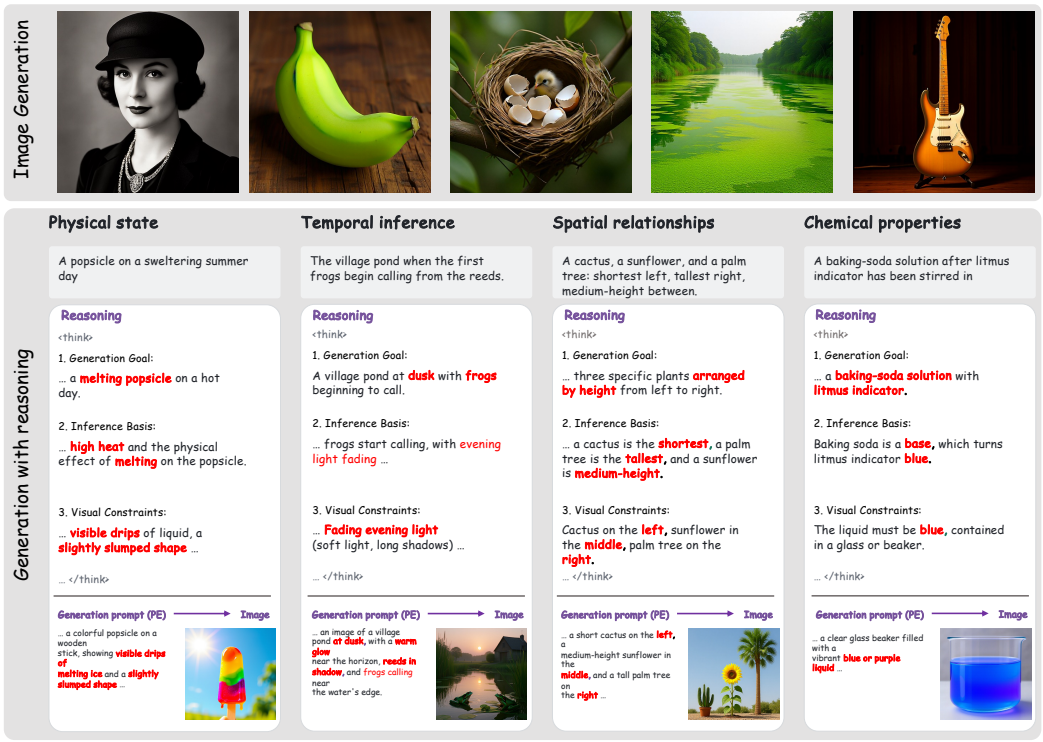


Figure 1: Qualitative examples of reasoning-driven image generation with DUETGRPO. Top: a gallery of generated images. Bottom: user instructions, structured reasoning, generation prompts (PE), and images across physical state, temporal inference, spatial relationships, and chemical properties. Highlighted text connects inferred requirements to visual constraints.

ing reasoning correctness and whether the generation prompt satisfies the user instruction, using task requirements as the evaluation criteria. We then examine how these text rewards should be combined with image–text alignment feedback from the evolving visual policy. Given the policies' distinct roles, we further ask whether shared reward feedback is sufficient for co-adaptation or whether cross-modality gradient coupling offers additional benefits. Our ablations support direct supervision of reasoning and prompt correctness and show further gains from sharing image–text alignment feedback with the language policy. Controls that retain the full reward favor continued updates to both policies over freezing either one, and detached conditioning over coupled gradients in the evaluated implementation. These findings motivate a design in which the language policy is rewarded for correct reasoning and prompts that satisfy the user instruction, shared image–text alignment feedback coordinates the two policies, and their gradients remain decoupled.

Guided by these findings, we propose DUETGRPO (Figure 2), a joint RL framework with complementary reasoning and image–text alignment rewards. A fixed text-only evaluator assesses reasoning correctness and checks the generation prompt against a candidate-independent checklist. This provides direct feedback on language outputs independently of rendering quality. A shared image–text alignment reward measures how faithfully each image follows its generation prompt. The language policy combines both rewards to learn prompts that satisfy the user instruction and can be rendered faithfully, while the visual policy uses only the alignment reward to improve prompt adherence. Joint online updates use group-relative advantages and decoupled gradients, so each policy follows its own objective while shared alignment feedback supports co-adaptation.

On the modular MLLM–DiT architecture, DUETGRPO improves the SFT initialization from 0.72 to 0.77 on WISE (Niu et al., 2025) and from 59.1 to 60.7 on T2I-CoReBench (Li et al., 2025). Applying the complete recipe to the mixture-of-transformers (MoT) model BAGEL (Deng et al., 2025) improves its SFT checkpoint from 0.64 to 0.66 on WISE (Niu et al., 2025).

Our contributions are threefold:

- **A joint RL framework for reasoning-driven image generation.** DUETGRPO directly rewards reasoning correctness and generation-prompt fidelity through a text-only evaluator, providing explicit supervision for the language policy. We combine this supervision with shared image–text alignment feedback to support online co-adaptation of language and visual policies.
- **Empirical insights into reward assignment and joint optimization.** We demonstrate the complementary benefits of direct text supervision and shared image–text alignment feedback and show the benefits of updating both policies with detached conditioning, supporting online co-adaptation through shared rewards and separate policy-gradient updates.
- **Effectiveness across benchmarks and architectures.** DuetGRPO improves WISE (Niu et al., 2025) Overall from 0.72 to 0.77 and T2I-CoReBench (Li et al., 2025) Overall from 59.1 to 60.7 on ThinkGen. Applying the same training recipe to MoTs further improves WISE (Niu et al., 2025) Overall from 0.64 to 0.66, demonstrating its applicability to both modular MLLM–DiT and MoT architectures.

2 RELATED WORK AND PRELIMINARIES

Unified models and reinforcement learning. Unified multimodal models either represent both modalities as discrete tokens (Wang et al., 2024; Geng et al., 2025; Chameleon Team, 2024; Wu et al., 2024) or combine language and visual policies in modular MLLM–DiT or mixture-of-transformers architectures (Deng et al., 2025; Chen et al., 2025a;b). Language-policy RL uses GRPO and reasoning supervision (Shao et al., 2024; Wei et al., 2022; Lightman et al., 2023; DeepSeek-AI et al., 2025), whereas visual-policy RL uses image–text alignment or human-preference rewards (Prabhudesai et al., 2023; Xu et al., 2024; Wallace et al., 2024; Black et al., 2023; Fan et al., 2023). Stochastic flow trajectories extend policy-gradient optimization to flow-matching visual policies (Liu et al., 2025; Xue et al., 2025). These image-based rewards do not directly evaluate reasoning correctness.

Joint RL for reasoning-driven image generation. For autoregressive generation, T2I-R1 jointly optimizes semantic-level planning and token-level image generation through BiCoT-GRPO with image-based rewards (Jiang et al., 2025). GoT-R1 (Duan et al., 2025) builds on Janus-Pro (Chen et al., 2025c), which autoregressively generates text and discrete image tokens. This differs substantially from our ThinkGen-based MLLM–DiT pipeline and BAGEL-based MoT implementation, which use diffusion-based visual generation. GoT-R1 combines prompt–reasoning, reasoning–image, and prompt–image consistency rewards into a shared sequence-level reward. Despite the architectural differences, its reward ablations support the importance of consistency rewards for joint RL of reasoning and generation. For MLLM–DiT pipelines, ThinkGen alternates policy updates (Jiao et al., 2025), while PromptRL, UniGRPO, and Think-Then-Generate study joint optimization (Wang et al., 2026b; Liu et al., 2026; Kou et al., 2026). We study how to assign complementary reasoning and image–text alignment rewards to the two policies during joint online training.

Credit assignment. DUETGRPO assigns text rewards specifically to the language policy and shares image–text alignment feedback across both policies during joint online updates (Appendix D.2, Figure 5). Image alignment and preference models evaluate rendered outputs (Radford et al., 2021; Wu et al., 2023b; Ma et al., 2025; Kirstain et al., 2023; Wu et al., 2023a; Wang et al., 2025; Lee et al., 2023). For open-ended language outputs, PromptEnhancer trains a dedicated reward model (Wang et al., 2026c); we use a fixed text-only evaluator to assess reasoning correctness and generation-prompt coverage and correctness against a candidate-independent checklist constructed from the instruction and its training reference explanation.

Problem formulation. For user instruction x, the language policy π_θ samples $z = (c, v)$, comprising a reasoning trace c and a generation prompt v, the text used to condition image generation. The visual policy p_ϕ then renders image y:

$$p_{\theta,\phi}(z, y \mid x) = \pi_\theta(z \mid x)\, p_\phi(y \mid v). \tag{1}$$

In the modular MLLM–DiT implementation, the SFT-initialized language policy is updated through LoRA (Hu et al., 2021) and the visual policy is optimized in full. Joint training updates θ and ϕ

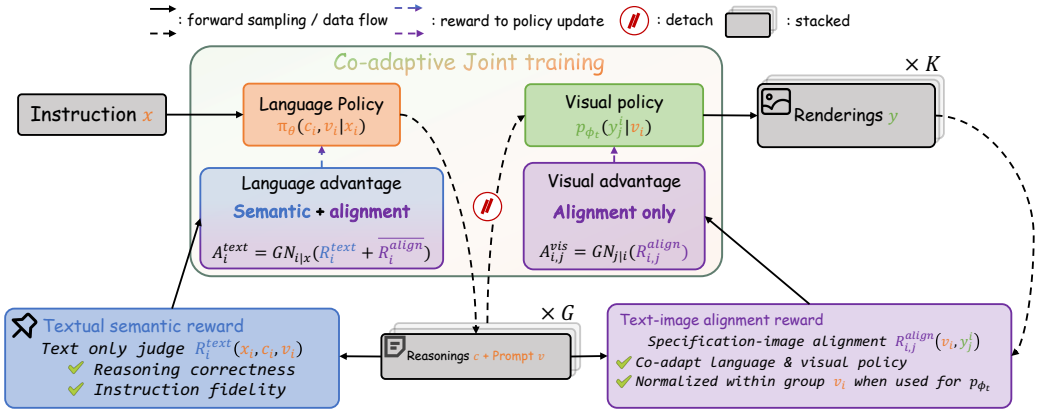


Figure 2: Overview of DUETGRPO. A text-only evaluator checks reasoning correctness and whether generation prompts satisfy the user instruction. The language policy combines this feedback with mean image–text alignment over its renderings; the visual policy uses alignment alone. Both policies update on fresh trajectories, while detached conditioning blocks gradients from the visual loss into the language policy.

from fresh trajectories within the same online loop; sharing a scalar image–text alignment reward does not require cross-modality gradient coupling.

GRPO objectives. For rewards $\{u_n\}_{n=1}^N$, let $\mathrm{GN}(u_n) = (u_n - \mu_u)/(\sigma_u + \zeta)$ denote group normalization, with group mean μ_u, standard deviation σ_u, and stabilizer $\zeta > 0$. Let $m \in \{\mathrm{lang}, \mathrm{vis}\}$ index the modality, with $\vartheta_{\mathrm{lang}} = \theta$, $\vartheta_{\mathrm{vis}} = \phi$, $\Omega_{\mathrm{lang}} = D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})$, and $\Omega_{\mathrm{vis}} = 0$. Defining $\mathcal{S}_\epsilon(\rho, A) = \min\{\rho A, \mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)A\}$, both objectives share the form (Schulman et al., 2017; Shao et al., 2024)

$$\mathcal{J}_m(\vartheta_m) = \mathbb{E}[\mathcal{S}_{\epsilon_m}(\rho^m, A^m) - \beta\Omega_m]\,. \tag{2}$$

For the language policy, the expectation is over outputs and tokens and ρ^{lang} is a token-probability ratio. For the visual policy, it is over renderings and denoising transitions; following FlowGRPO, ρ^{vis} is a stochastic transition ratio (Liu et al., 2025). We construct the two advantages to combine direct language supervision with shared rendering feedback at their respective comparison levels.

3 METHODOLOGY

DUETGRPO combines composite textual supervision with shared image–text alignment feedback. Both policies update online, while detached conditioning keeps their policy gradients decoupled.

3.1 BASE MODEL AND SUPERVISED INITIALIZATION

We use ThinkGen (Jiao et al., 2025) as our primary base model, pairing a Qwen3-VL-8B-based MLLM (Bai et al., 2025) with an OmniGen2-derived 4B DiT (Wu et al., 2025b). The MLLM autoregressively produces a reasoning trace c followed by a generation prompt v. ThinkGen's visual generation instruction refinement (VGI-refine) extracts the prompt-token hidden states from the final two MLLM layers and combines them with learnable prepadding states. A linear connector maps these representations into conditioning for the DiT, which synthesizes image latents decoded by a VAE. In Eq. (1), $p_\phi(y \mid v)$ abbreviates rendering conditioned on these MLLM-derived representations, which evolve with the language policy.

Before joint RL, we apply SFT to the language policy on a mixture of existing caption corpora and newly synthesized knowledge-dependent captions. Gemini 3.1 Pro Preview annotates every source with the same structured CoT and prompt-enhancement (PE) target, teaching the MLLM to translate an instruction into a reasoned, renderable description. The resulting checkpoint is denoted DUETGRPO (SFT); Section 4.1 summarizes the training sets, and Appendix A details their composition and response format.

Starting from this initialization, each online iteration samples G outputs $z_i = (c_i, v_i) \sim \pi_\theta(\cdot \mid x)$ per user instruction and K images $y_{ij} \sim p_\phi(\cdot \mid v_i)$ per generation prompt. Both policies update before the next trajectories are collected. Appendix C provides the training recipe, including rollout and optimization hyperparameters.

3.2 Composite Text Reward

For each instruction x, a fixed text-only evaluator constructs and caches a candidate-independent checklist $\mathcal{C}(x)$ specifying the required subject, outcome, and supporting visual attributes and relations. The instruction's training reference explanation supplies the factual basis for this checklist. All candidates share the same evaluation criteria, accommodating different valid reasoning paths and generation prompts.

For each output $z_i = (c_i, v_i)$, the evaluator assigns three scores: reasoning correctness $r_i \in \{0, \frac{1}{2}, 1\}$ for the conclusion and its factual support, prompt coverage $p_i \in [0, 1]$ for the checklist requirements expressed in v_i, and prompt correctness $h_i \in \{0, 1\}$ for the validity of its depicted subject and outcome. We combine them as

$$R_i^{\text{text}} = \frac{1}{5} r_i + \frac{3}{5} p_i + \frac{1}{2} h_i. \tag{3}$$

The language policy combines this reward with mean image–text alignment feedback (Section 3.3). Appendix B details the scoring rules and component ablations.

3.3 Reward Routing and Joint Optimization

We complement the text reward with CLIPScore (Radford et al., 2021), $R_{ij}^{\text{align}} = V_{\text{CLIP}}(y_{ij}, v_i)$, which measures image alignment with the generation prompt, excluding the reasoning trace. Let $\overline{R}_i^{\text{align}}$ denote its mean over the K renderings of v_i. The language policy receives both rewards, while the visual policy receives alignment alone. Their group-relative reward signals are:

$$\widetilde{A}_i^{\text{lang}} = \text{GN}_{i=1}^{G}\left(R_i^{\text{text}} + \overline{R}_i^{\text{align}}\right), \qquad \widetilde{A}_{ij}^{\text{vis}} = \text{GN}_{j=1}^{K}\left(R_{ij}^{\text{align}}\right). \tag{4}$$

The raw language rewards are summed with equal weights before normalization across candidates for the same instruction. Averaging alignment over K renderings estimates how well the current visual policy follows each candidate prompt, reducing dependence on any single sampled image. Together with the composite text reward, this feedback encourages prompts that satisfy the user instruction and can be rendered faithfully. Within-prompt centering makes the visual advantage invariant to prompt-level text scores and rewards each trajectory for its relative alignment under the same condition.

To obtain the optimization advantages A^{lang} and A^{vis} in Eq. (2), the implementation uses a standard-deviation floor for language normalization and applies batch standardization and clipping to $[-5, 5]$ to the visual signals (Appendix C). Each image's visual advantage weights the clipped likelihood-ratio objective at every selected stochastic denoising transition along its sampled trajectory. Rewards and sampled latents are treated as fixed, and language-derived conditioning is detached.

Both policies update before fresh trajectories are collected. The language policy uses alignment feedback from the evolving renderer, and the visual policy trains on prompts from the evolving language policy. This online co-adaptation uses shared scalar feedback with detached conditioning.

4 Experiments

We evaluate DuetGRPO on WISE (Niu et al., 2025) and T2I-CoReBench (Li et al., 2025), examine text supervision with a frozen visual policy, and ablate text rewards, shared alignment, policy freezing, and gradient routing on ThinkGen. We also apply the complete recipe to BAGEL.

4.1 Evaluation Setup

Training data. Our additional SFT corpus contains approximately 258K examples from BLIP3o (Chen et al., 2025a), T2I80K, FLUX-Reason-6M (Fang et al., 2025), and synthetic knowledge captions, all annotated with structured CoT+PE by Gemini 3.1 Pro Preview. Joint RL uses approximately 6.2K requests: 2K filtered from ThinkGen's reasoning-generation pool (Jiao et al., 2025) and 4.2K reused from our synthetic SFT captions. Appendix A details data composition and annotation.

Comparison methods. We implement ThinkGen adaptations of PromptRL (Wang et al., 2026b), UniGRPO (Liu et al., 2026), and Dual-GRPO (Think-Then-Generate) (Kou et al., 2026), sharing the backbone, SFT initialization, training data, and inference/evaluation protocol with DuetGRPO. Dual-GRPO does not specify which model implements its semantic-consistency reward, while the RewardDance (Wu et al., 2025c) reward model used by UniGRPO has not been publicly released. We therefore use CLIPScore as a common image–text consistency scorer for these two adaptations and DuetGRPO. UniGRPO and Dual-GRPO retain their original reward inputs, comparing the user instruction x with image y through $V_{\mathrm{CLIP}}(y, x)$. For DuetGRPO, alignment compares the image with generation prompt v, $R^{\mathrm{align}} = V_{\mathrm{CLIP}}(y, v)$, while R^{text} assesses reasoning and prompt fidelity against x. The PromptRL adaptation retains its original prompt-retention mechanism, grouping images from original and refined prompts for advantage estimation. These results compare the implemented ThinkGen configurations under our common training setup. Table 4 and Appendices C and D.3 specify their rewards, gradient routing, and control definitions.

Benchmarks. For WISE (Niu et al., 2025), we use Qwen3.5-35B-A3B as the judge following the official evaluation protocol (PKU-YuanGroup, 2026), and report category scores and category-weighted Overall WiScore on $[0, 1]$. The same WISE protocol is used for the BAGEL experiment, text-reward study, and joint-training comparisons in Tables 3 and 4. For T2I-CoReBench (Li et al., 2025), we use Gemini 2.5 Flash and report scores on $[0, 100]$. Table 1 includes all 12 dimensions and Composition, Reasoning, and Overall aggregates for the selected baselines from the official leaderboard, alongside the available aggregate scores for ThinkGen and our models. Sampling and aggregation are detailed in Appendix D.1.

4.2 Experimental Results

We report T2I-CoReBench (Li et al., 2025) and WISE (Niu et al., 2025) separately in Tables 1 and 2. The T2I-CoReBench table reports all 12 dimensions and aggregate scores for the selected baselines under Gemini 2.5 Flash evaluation; the WISE table reports six categories and Overall. The SFT–GRPO comparison measures the effect of our training recipe; published reference scores provide context under their respective inference configurations.

Benchmark comparison. Table 1 shows that DuetGRPO raises T2I-CoReBench (Li et al., 2025) Overall from 59.1 to 60.7. Reasoning improves by 1.9 points ($59.6 \rightarrow 61.5$), compared with 1.0 for composition ($58.1 \rightarrow 59.1$), consistent with the goal of translating inferred requirements into visual content. In Table 2, WISE (Niu et al., 2025) Overall rises from 0.72 to 0.77, with the largest gains in Physics ($+0.11$), Space ($+0.09$), and Chemistry ($+0.08$). The following controls examine which parts of the recipe support these generation gains.

Table 3(b) evaluates composite text supervision with a frozen visual policy. Panel (a) reports the application to BAGEL, discussed in Section 4.4.

4.2.1 Effectiveness of Composite Textual Supervision

With the visual policy frozen, updating language using R^{text} raises Overall from 0.72 to 0.74 and Physics from 0.61 to 0.70 (Table 3b), demonstrating that direct text supervision improves the generation prompts supplied to the renderer. The component ablations in Appendix B.2 show that the complete reward outperforms every component ablation in Overall and all three reported categories, supporting the complementary contributions of reasoning correctness, prompt coverage, and prompt correctness. Section 4.3 examines how this supervision combines with shared image–text alignment feedback during joint training.

Table 1: T2I-CoReBench (Li et al., 2025) results with Gemini 2.5 Flash as the evaluator ($[0, 100]$, higher is better). For the selected baselines, we report all 12 dimensions, both capability means, and Overall from the official leaderboard (T2I-CoReBench Team, 2026) (accessed September 26, 2026). ThinkGen and our models retain their previously reported aggregate scores; dashes denote unreported scores. Overall averages the 12 dimensions.

Model	**Composition**					**Reasoning**									**Overall**
	MI	MA	MR	TR	**Mean**	LR	BR	HR	PR	GR	AR	CR	RR	**Mean**	↑
FLUX.2-klein-4B (Black Forest Labs, 2025)	80.0	77.1	63.5	53.0	68.4	34.6	30.5	49.5	77.3	58.1	58.1	48.8	25.2	47.7	54.6
FLUX.2-klein-9B (Black Forest Labs, 2025)	85.8	80.6	68.0	77.7	78.0	43.4	32.9	52.4	80.6	60.2	60.8	60.4	25.1	52.0	60.6
LongCat-Image (LongCat Team, 2025)	81.4	74.5	61.5	65.7	70.8	39.1	35.7	48.5	75.5	72.5	61.4	58.8	41.0	54.1	59.6
Z-Image-Turbo (Z-Image Team et al., 2025)	79.5	72.2	62.7	83.9	74.6	36.9	28.8	48.7	74.0	56.2	55.8	52.0	26.0	47.3	56.4
FLUX.1-Krea-dev (T2I-CoReBench Team, 2026)	70.7	71.1	53.2	28.9	56.0	30.3	26.1	44.5	70.6	50.5	57.5	46.3	28.7	44.3	48.2
BAGEL w/ Think (Deng et al., 2025)	57.7	60.8	37.8	2.2	39.6	25.5	25.4	33.9	58.6	53.5	56.9	41.6	39.8	41.9	41.1
show-o2-7B (T2I-CoReBench Team, 2026)	59.4	61.8	38.1	2.2	40.4	23.2	23.1	37.5	51.6	40.9	47.2	32.2	21.3	34.6	36.5
FLUX.1-dev (Black Forest Labs, 2024)	58.6	60.3	44.1	31.1	48.6	24.8	23.0	36.0	61.8	42.4	57.2	36.3	30.3	39.0	42.2
HunyuanImage 3.0 (Cao et al., 2025)	84.9	81.2	63.7	85.7	78.9	39.6	32.8	51.4	72.4	54.1	54.1	57.0	27.7	48.6	58.7
Qwen-Image (Wu et al., 2025a)	81.4	79.6	65.6	85.5	78.0	41.1	32.2	48.2	75.1	56.5	53.3	61.9	26.4	49.3	58.9
Janus-Pro-7B (Chen et al., 2025c)	54.4	59.3	40.9	7.5	40.5	19.8	20.9	34.6	22.4	11.5	30.4	8.7	9.8	19.8	26.7
BLIP3-o-8B (Chen et al., 2025a)	46.2	50.4	24.1	0.5	30.3	14.8	20.7	28.3	39.6	43.4	51.0	35.9	20.4	31.8	31.3
OmniGen2-7B (Wu et al., 2025b)	67.9	64.1	48.3	19.2	49.9	24.7	23.2	43.3	63.1	46.1	54.2	36.5	24.1	39.4	42.9
ThinkGen w/ CoT (Jiao et al., 2025)	–	–	–	–	60.1	–	–	–	–	–	–	–	–	58.9	59.3
DUETGRPO (SFT)	73.0	69.1	52.4	38.0	58.1	48.3	45.2	55.5	71.7	72.4	60.5	54.8	68.5	59.6	59.1
DUETGRPO (GRPO)	77.5	72.8	55.9	30.2	59.1	46.2	49.4	57.6	73.0	72.6	62.3	59.7	72.0	61.5	**60.7**

Table 2: WISE (Niu et al., 2025) benchmark comparison ($[0, 1]$, higher is better). Our evaluations use Qwen3.5-35B-A3B; published reference scores follow the cited evaluations and leaderboard (PKU-YuanGroup, 2026). Overall is category-weighted. Dashes denote unavailable category scores. See Appendices D.1 and D.3 for evaluation protocols and reference sources.

Model	**Cultural**	**Time**	**Space**	**Biology**	**Physics**	**Chemistry**	**Overall ↑**
BAGEL w/ Think (Deng et al., 2025)	0.78	0.63	0.56	0.37	0.55	0.50	0.63
FLUX.1-dev (Black Forest Labs, 2024)	0.52	0.40	0.53	0.17	0.37	0.24	0.41
HunyuanImage 3.0 (Cao et al., 2025)	0.53	0.39	0.48	0.31	0.45	0.24	0.44
Qwen-Image (Wu et al., 2025a)	0.63	0.53	0.56	0.34	0.48	0.25	0.51
Janus-Pro-7B (Chen et al., 2025c)	0.30	0.37	0.49	0.36	0.42	0.26	0.35
BLIP3-o-8B (Chen et al., 2025a)	–	–	–	–	–	–	0.62
OmniGen2-7B (Wu et al., 2025b)	0.42	0.52	0.64	0.43	0.50	0.34	0.47
ThinkGen w/ CoT (Jiao et al., 2025)	0.81	0.67	0.75	0.53	0.64	0.64	0.71
SDXL (Podell et al., 2023)	0.43	0.48	0.47	0.44	0.45	0.27	0.43
Z-Image (Z-Image Team et al., 2025)	0.55	0.47	0.51	0.33	0.48	0.18	0.45
Emu3-Gen (Wang et al., 2024)	0.34	0.45	0.48	0.41	0.45	0.27	0.39
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
DeepGen 1.0 (Wang et al., 2026a)	0.65	0.41	0.72	0.39	0.59	0.45	0.57
DUETGRPO (SFT)	0.79	0.77	0.71	0.62	0.61	0.66	0.72
DUETGRPO (GRPO)	0.81	0.82	0.80	0.62	0.72	0.74	**0.77**

The paired examples in Appendix E show the final prompt restoring the origami paper's purple color and changing the Delhi scene from night to late afternoon. These examples illustrate how text supervision improves the generation prompts supplied to a frozen visual policy.

4.3 REWARD AND GRADIENT-ROUTING CONFIGURATIONS

Table 4 examines the contributions of reward assignment, policy adaptation, and gradient routing on ThinkGen with shared data and initialization. Entries with $\pm$ summarize the mean and standard deviation of final WISE Overall scores across three training seeds under the same budget. Figure 3 traces image-reward optimization within each method: CLIPScore measures alignment with user instructions for UniGRPO and Dual-GRPO, and with generation prompts for DUETGRPO.

Table 3: Composite text supervision and application to BAGEL on WISE (Niu et al., 2025), evaluated with Qwen3.5-35B-A3B. (a) applies the same SFT and complete DUETGRPO joint-training recipe to BAGEL. (b) evaluates the composite text reward on the ThinkGen-based backbone, with the visual policy frozen and no image-based reward during GRPO. "Original" denotes each study's starting checkpoint. Bold marks the best score within each panel, including ties.

(a) BAGEL: complete DUETGRPO

Stage	Cultural	Physics	Chemistry	Overall ↑
Original	0.78	0.55	0.50	0.63
SFT	**0.79**	0.57	0.63	0.64
SFT + DUETGRPO	0.78	**0.61**	**0.66**	**0.66**

(b) ThinkGen: text-only reward

Stage	Cultural	Physics	Chemistry	Overall ↑
Original	0.77	0.56	0.61	0.65
SFT	0.79	0.61	0.66	0.72
SFT + GRPO	**0.80**	**0.70**	**0.67**	**0.74**

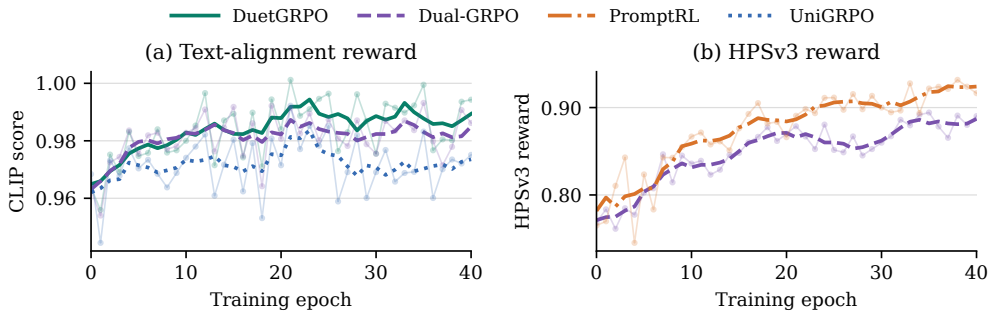


Figure 3: Training rewards on ThinkGen: (a) CLIPScore; (b) HPSv3. Faint curves show epoch means; solid curves show centered five-epoch moving averages, truncated at boundaries.

Joint-training gains. DUETGRPO achieves the highest WISE (Niu et al., 2025) Overall in Table 4, improving the shared initialization from 0.72 to 0.77 and exceeding the strongest adapted baseline, Dual-GRPO (0.74), by 0.03. Figure 4 complements these quantitative gains with examples spanning common knowledge identification, animal depth ordering, desert-tortoise shelter, chameleon camouflage, and litmus color, illustrating how instruction requirements are expressed in generated images.

Complementary supervision and alignment feedback. With both policies trainable and conditioning detached, adding R^{text} to the alignment-based language reward raises WISE Overall from 0.75 to 0.77. Adding mean image–text alignment feedback to the text-based language reward yields the same 0.02 gain, with the visual alignment reward fixed in both comparisons. These paired gains establish the complementary roles of the two signals: text supervision evaluates reasoning and prompt fidelity against the user instruction, while shared alignment feedback rewards prompts that the evolving renderer follows faithfully. Their combination connects instruction-level correctness with rendering performance throughout joint training.

Co-adaptation and gradient routing. Under the full language reward $R^{\text{text}} + \overline{R}^{\text{align}}$, updating both policies achieves 0.77 WISE Overall, improving on language-only updates (0.73) and vision-only updates (0.74). With both policies trainable, detached conditioning further improves Overall from 0.75 with coupled gradients to 0.77. Together, these results support online co-adaptation through shared rewards and separate policy-gradient updates: each policy learns from the evolving counterpart while following its assigned objective. Appendix F.1 complements these comparisons with an analysis of how conditioning gradients shape the frequency and composition of language-parameter updates.

4.4 APPLICATION TO BAGEL

On the mixture-of-transformers model BAGEL (Deng et al., 2025), WISE Overall under Qwen3.5-35B-A3B evaluation increases from 0.63 to 0.64 after SFT and to 0.66 after DUETGRPO (Table 3a). The gain over SFT supports applying the complete recipe to BAGEL's MoT architecture.

Table 4: Joint-training comparisons and ablations on ThinkGen with shared data and initialization. The first three methods are adapted baselines. R^{text} scores reasoning and prompt correctness; overlines denote means over renderings of the same prompt. CLIPScore uses user instructions for UniGRPO and Dual-GRPO, and generation prompts otherwise. WISE (Niu et al., 2025) Overall ($[0, 1]$) entries with $\pm$ report mean $\pm$ standard deviation across three runs with different training seeds and the same training budget; Δ is mean Overall minus Initial. See Appendix D.3 for control and evaluation details.

Training variant	Language reward	Visual reward	Gradients	Initial	Overall ↑	Δ
PromptRL (Wang et al., 2026b)	HPSv3	HPSv3	Detached	0.72	0.69	−0.03
UniGRPO (Liu et al., 2026)	CLIPScore	CLIPScore	Coupled	0.72	0.73	+0.01
Dual-GRPO (Kou et al., 2026)	CLIPScore	CLIPScore + HPSv3	Detached	0.72	0.74	+0.02
Language without text reward	$\overline{\text{CLIPScore}}$	CLIPScore	Detached	0.72	0.75 $\pm$0.0077	+0.03
Language without alignment	R^{text}	CLIPScore	Detached	0.72	0.75 $\pm$0.0055	+0.03
Coupled	$R^{\text{text}} + \overline{\text{CLIPScore}}$	CLIPScore	Coupled	0.72	0.75 $\pm$0.0055	+0.03
Freeze vision	$R^{\text{text}} + \overline{\text{CLIPScore}}$	Frozen	Detached	0.72	0.73 $\pm$0.0014	+0.01
Freeze language	Frozen	CLIPScore	Detached	0.72	0.74 $\pm$0.0071	+0.02
DUETGRPO (GRPO)	$R^{\text{text}} + \overline{\text{CLIPScore}}$	CLIPScore	Detached	0.72	**0.77** $\pm$0.0063	**+0.05**

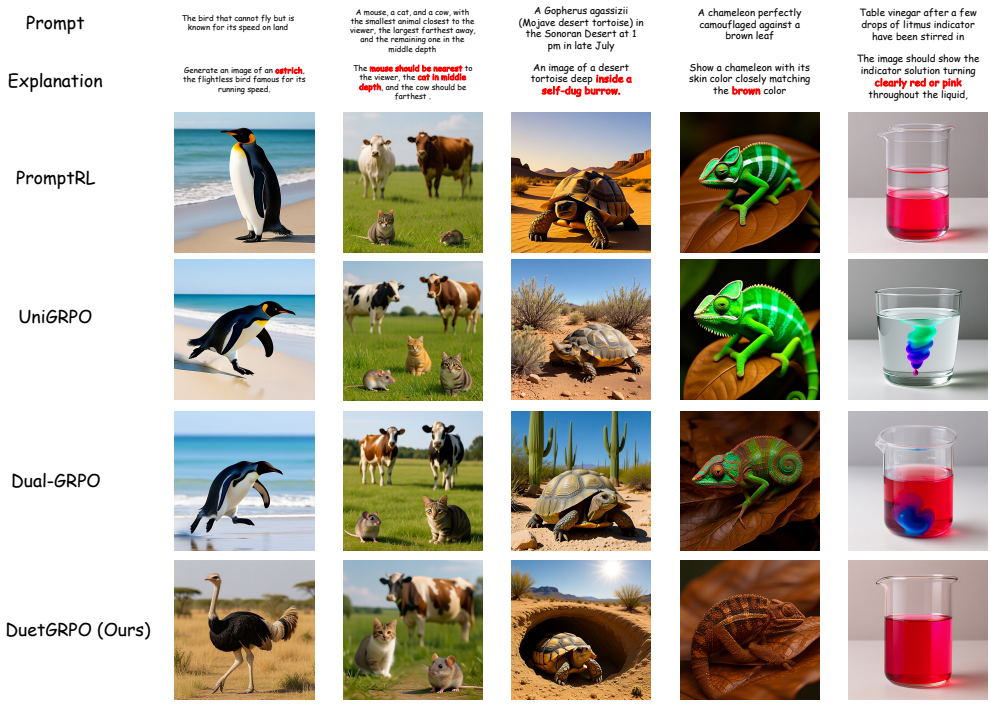


Figure 4: Selected WISE (Niu et al., 2025) outputs by different methods.

5 CONCLUSION

We presented DUETGRPO, a joint RL recipe that combines composite textual supervision with shared image–text alignment feedback. Both policies update from fresh online trajectories, and detached conditioning blocks visual-loss gradients into language while preserving shared scalar feedback. Improvements over the same SFT initialization on WISE (Niu et al., 2025) and T2I-CoReBench (Li et al., 2025) support the complete recipe. ThinkGen ablations show complementary gains from direct text supervision and shared alignment feedback under joint updates, while further controls favor updating both policies with detached conditioning. The same recipe improves BAGEL, extending its application to a MoT architecture.

AI USE STATEMENT

We used Gemini models to generate synthetic captions and Gemini 3.1 Pro Preview to annotate reasoning traces and generation prompts and compute text rewards. Gemini 2.5 Flash and Qwen3.5-35B-A3B served as benchmark judges. These uses are described in Section 4.1 and Appendices A, B, and D.1. We also used OpenAI Codex to assist with manuscript revision, language editing, and LaTeX preparation. The authors retain full responsibility for the final manuscript, experimental results, and associated artifacts, including AI-assisted content.

ETHICS STATEMENT

DUETGRPO aims to improve instruction following in text-to-image generation using existing datasets, synthetic captions, and automated feedback. More faithful generation can also increase the risk of misleading imagery, impersonation, and harmful content. Biases and factual errors in pretrained models, source data, synthetic annotations, and model-based rewards may be inherited or amplified during reinforcement learning. Our benchmark evaluations assess generation quality and reasoning, and do not establish fairness, factual reliability, or safety for deployment. Downstream use should include content safeguards, clear disclosure of synthetic imagery, and respect for data and model licenses, privacy, and consent.

REPRODUCIBILITY STATEMENT

Section 3 describes model initialization, policy objectives, reward assignment, and gradient routing. Appendix A documents the data sources, subset sizes, annotation procedure, and response format; Appendix B specifies the composite reward and scoring rules. Appendix C and Table 7 report rollout, optimization, and inference settings. Benchmark protocols and score aggregation are given in Appendix D.1, while baseline adaptations, controls, and repeated-run aggregation are detailed in Appendix D.3. The separate gradient-probe configurations are described in Appendix F.1. Exact replication depends on access to the internal T2I80K data and the specified model services, as well as the complete reward-scoring system prompts planned for release in Appendix B.

REFERENCES

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.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

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.

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 preprint arXiv:2509.23951*, 2025.

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

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, Tianyi Zhou, Junnan Li, Silvio Savarese, Caiming Xiong, and Ran Xu. BLIP3o-NEXT: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025c.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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 preprint 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 preprint arXiv:2505.17022*, 2025. URL https://arxiv.org/abs/2505.17022.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-Omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. *arXiv preprint arXiv:2505.00703*, 2025. URL https://arxiv.org/abs/2505.00703.

Siyu Jiao, Yiheng Lin, Yujie Zhong, Qi She, Wei Zhou, Xiaohan Lan, Zilong Huang, Fei Yu, Yingchen Yu, Yunqing Zhao, et al. ThinkGen: Generalized thinking for visual generation. *arXiv preprint arXiv:2512.23568*, 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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023.

Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yugang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, et al. Think-Then-Generate: Reasoning-aware text-to-image diffusion with LLM encoders. *arXiv preprint arXiv:2601.10332*, 2026.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? *arXiv preprint arXiv:2509.03516*, 2025.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

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 reinforcement learning. *arXiv preprint arXiv:2505.05470*, 2025.

Jie Liu, Zilyu Ye, Linxiao Yuan, Shenhan Zhu, Yu Gao, Jie Wu, Kunchang Li, Xionghui Wang, Xiaonan Nie, Weilin Huang, and Wanli Ouyang. UniGRPO: Unified policy optimization for reasoning-driven visual generation. *arXiv preprint arXiv:2603.23500*, 2026.

LongCat Team. LongCat-Image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Chaoran Feng, Bin Zhu, and Li Yuan. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

PKU-YuanGroup. WISE leaderboard: Qwen3.5-35B verified batch results. `https://github.com/PKU-YuanGroup/WISE/blob/main/leadboard.md`, 2026. Accessed July 28, 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.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

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

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

T2I-CoReBench Team. T2I-CoReBench leaderboard. `https://t2i-corebench.github.io/#leaderboard`, 2026. Gemini 2.5 Flash evaluation results; accessed September 26, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

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 preprint arXiv:2602.12205*, 2026a.

Fu-Yun Wang, Han Zhang, Michael Gharbi, Hongsheng Li, and Taesung Park. PromptRL: Prompt matters in RL for flow-based image generation. *arXiv preprint arXiv:2602.01382*, 2026b.

Linqing Wang, Zhiyong Xu, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Zhenxi Li, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. PromptEnhancer: Taming your rewriter for text-to-image generation via fine-grained reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14895–14904, 2026c.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

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

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

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 preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. RewardDance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025c.

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

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023b.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. VILA-U: A unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

A Data Composition and CoT Format

Supervised data sources. Table 5 separates our additional SFT data from the subsequent RL request pool. Our SFT mixture combines the BLIP3o (Chen et al., 2025a) and T2I80K source subsets, a random 30K-example subset of FLUX-Reason-6M (Fang et al., 2025), and 89.1K newly generated captions. T2I80K is our local name for an 80K-example subset sampled from an internal image–text collection. The synthetic captions cover general chemistry, physics, spatiotemporal reasoning, biology, and general world knowledge. This mixture supplies both general visual descriptions and requests whose visual realization requires knowledge or inference.

Table 5: Data used for our supervised adaptation and joint RL. BLIP3o (Chen et al., 2025a) and T2I80K counts are from the source-subset manifests; K denotes rounded thousands for the other components. Counts refer to selected records.

Stage	Source	Examples	Construction
SFT	BLIP3o (Chen et al., 2025a)	58,854	Source subset
	T2I80K	79,993	Internal sample
	FLUX-Reason-6M (Fang et al., 2025)	30K	Random sample
	Synthetic knowledge captions	89.1K	Gemini-generated captions
	Total	≈258K	CoT+PE for every source
RL	ThinkGen reasoning-generation pool (Jiao et al., 2025)	2K	Filtered from 10K
	Synthetic knowledge captions	4.2K	Independent of the SFT pool
	Total	≈6.2K	Requests for online rollouts

Unified CoT and prompt-enhancement supervision. Caption acquisition and response annotation are separate steps: only the 89.1K-caption component is newly synthesized, whereas Gemini 3.1 Pro Preview annotates *all four* SFT sources with our prescribed response format. For each input instruction x, it generates a structured reasoning trace c and an enhanced generation prompt v (PE). Thus, every source supervises the same instruction-to-CoT-to-prompt interface, rather than introducing a source-specific output format. The five-field template below defines the reasoning structure, and the text following the reasoning block provides the final generation prompt.

RL requests and cross-stage reuse. The RL pool combines approximately 2K requests filtered from ThinkGen's 10K reasoning-generation dataset (Jiao et al., 2025) with 4.2K synthetic captions independent of our synthetic SFT corpus. For each request, training samples G language candidates and K images per candidate online and optimizes the resulting trajectories using the rewards in Section 3.

CoT format. We prompt Gemini 3.1 Pro Preview to organize each reasoning trace into five fields: (1) *Generation Goal*, which states the intended visual outcome; (2) *Inference Basis*, which records the knowledge or relations needed to reason about the user instruction; (3) *Visual Constraints*, which identifies the evidence that must appear in the image; (4) *Composition Plan*, which translates that evidence into a scene layout; and (5) *Final Instruction Check*, which verifies that the resulting instruction is complete and renderable. The exact response template is

```
[SYS] [C] <think>
1. Generation Goal: [goal]
2. Inference Basis: [reasoning]
3. Visual Constraints: [constraints]
4. Composition Plan: [composition]
5. Final Instruction Check: [verification]
</think>
[final generation prompt]
```

where `[C]` is the user instruction. Only the text after the closing `</think>` token is passed to the renderer, while the complete trace provides supervision for the language-side reasoning process.

B Composite Text Reward Details

We use a fixed Gemini 3.1 Pro Preview evaluator. For each training instruction x, the evaluator uses its reference explanation to resolve the required subject and outcome and caches a shared checklist $\mathcal{C}(x)$. The checklist contains at most two critical requirements—the subject and, when applicable, the tested outcome—with supporting requirements for visible attributes and relations. Each candidate $z_i = (c_i, v_i)$ is assessed through the following three scores.

Reasoning correctness, r_i. The evaluator assigns $r_i = 1$ when c_i reaches the correct conclusion and supports it with a sound factual basis, such as a relevant property, mechanism, or convention. It assigns $r_i = \frac{1}{2}$ when the conclusion is correct but its justification is incomplete or partly flawed, and $r_i = 0$ when the conclusion is incorrect or unresolved. For instructions that require visual planning, a sound plan provides the supporting basis.

Prompt coverage, p_i. Each checklist requirement receives credit 1, $\frac{1}{2}$, or 0 according to whether v_i expresses it fully, partially, or leaves it unsatisfied. Full credit requires a concrete subject or a visually explicit description of the required state, attribute, or relation; partial credit captures a valid but underspecified realization. Compatible choices of style, setting, and composition are accepted.

We compute an overall coverage fraction and a supporting-requirement coverage fraction by dividing the corresponding credit sum by the number of requirements. Each added description that contradicts the task, introduces a competing subject or state, or obscures the required subject contributes one zero-credit entry to both denominators, up to four entries. The supporting fraction is 1 when its denominator is zero. The following rules determine p_i, in order:

1. A factually incompatible subject receives $p_i = 0$.
2. When the checklist contains no critical requirement, p_i equals the supporting coverage fraction.
3. When every critical requirement receives full credit, p_i is 0.55 plus 0.45 times the supporting coverage fraction.
4. In the remaining cases, p_i is 0.55 times the overall coverage fraction.

Prompt correctness, h_i. The evaluator assigns $h_i = 1$ when v_i commits to the intended subject and expresses a compatible outcome in recognizable visual terms. It assigns $h_i = 0$ for a wrong subject, an incompatible or competing outcome, or a description that leaves the required subject or outcome visually unidentifiable. The evaluator considers c_i together with v_i to resolve the intended meaning and accepts semantically equivalent visual realizations. Supporting attributes contribute through p_i, allowing a semantically correct prompt to receive the correctness bonus while retaining partial coverage credit. The three scores enter the additive reward in Eq. (3).

B.1 Illustrative Checklist

The following illustrative example shows a simple identification task. Since the instruction asks for an entity without testing a state change, only the subject is tagged `critical`; its recognizable visual attributes are tagged `minor`.

> **User prompt:** The round pastry shared during Mid-Autumn Festival.

An example `checklist` contains the following cues:

Cue	Tag
A mooncake	`critical`
Round shape	`minor`
Golden-brown baked crust	`minor`
Decorative stamped pattern on the top surface	`minor`

The same checklist applies to every candidate for this instruction. The evaluator subsequently checks whether each candidate generation prompt expresses these cues, while assessing the reasoning trace

separately. We will open-source the complete set of system prompts used for reward scoring in a future release.

B.2 Component Ablations

We isolate the contributions of reasoning correctness r_i, prompt coverage p_i, and prompt correctness h_i by ablating each term in Eq. (3) under the frozen-visual-policy, text-only GRPO setting of Table 3(b). Table 6 compares these objectives with the SFT initialization and the full composite reward.

Table 6: Composite text reward ablations on WISE (Niu et al., 2025) with a frozen visual policy. Overall aggregates all six categories. The full reward achieves the highest score in every reported column; bold marks column maxima.

Variant	Cultural	Physics	Chemistry	Overall ↑
ThinkGen SFT (no RL)	0.79	0.61	0.66	0.72
Without r_i (reasoning)	0.79	0.60	0.66	0.73
Without p_i (coverage)	0.79	0.63	0.60	0.72
Without h_i (correctness)	0.78	0.57	0.64	0.72
Full composite reward	**0.80**	**0.70**	**0.67**	**0.74**

The full composite reward achieves the highest score in every reported metric, reaching 0.74 Overall, 0.80 Cultural, 0.70 Physics, and 0.67 Chemistry. Each component contributes to aggregate performance: adding reasoning correctness to the other two terms raises Overall from 0.73 to 0.74, while adding either prompt coverage or prompt correctness to its corresponding two-term objective raises Overall from 0.72 to 0.74.

The category-level gains further reveal the value of these complementary criteria. Reasoning supervision raises Physics from 0.60 to 0.70; prompt coverage improves both Physics and Chemistry by 0.07; and prompt correctness raises Physics from 0.57 to 0.70. Together, these results support supervising the full reasoning-to-prompt process: a sound factual basis, explicit coverage of the required visual content, and a correct depicted subject and outcome. Combining all three criteria yields the strongest generation performance across the reported metrics.

C Training Recipe

Table 7 records the implementation settings of the ThinkGen-based DuetGRPO run whose CLIP-Score trajectory appears in Figure 3(a).

Initialization and trainable parameters. We initialize the ThinkGen-based model from ThinkGen's publicly released SFT checkpoint before our additional SFT, so the "Original" scores in Table 3(b) differ from the published ThinkGen results in Table 2. Joint RL starts from the supervised checkpoint described in Section 3.1, using the response format in Appendix A. We train LoRA (Hu et al., 2021) adapters in the language policy and all parameters of the DiT. The policies have separate optimizers, and the language-derived conditioning is cached during rollout and detached for visual updates. Training uses bfloat16 mixed precision and gradient checkpointing.

Online rollout collection. For each user instruction, we sample $G = 4$ reasoning traces and generation prompts at temperature 1.0, with a maximum of 1,024 generated tokens per response. Each generation prompt conditions $K = 6$ image trajectories, yielding 24 images per instruction. The run shares the initial latent across this candidate–image grid and uses stochastic Flow-SDE trajectories (Liu et al., 2025; Xue et al., 2025). Each trajectory has 30 denoising steps; three stochastic steps are sampled from scheduler indices $\{1, 2, 3, 4, 5\}$ with noise level 0.6. Training uses 512×512 images and guidance scale (Ho & Salimans, 2022) 4.0. An online epoch collects five rollout batches and then updates both policies; it is a collection–update cycle rather than a full pass over the dataset.

Reward computation and normalization. The frozen text-only evaluator is Gemini 3.1 Pro Preview, and the alignment scorer is CLIP ViT-L/14 (Radford et al., 2021). The language reward adds the raw composite text reward to mean CLIPScore over six images, both with coefficient 1. We center this sum within each four-candidate group and divide by $\max(\sigma_x, 0.3\sigma_{\text{batch}}, 10^{-6})$, where σ_x is the group standard deviation and σ_{batch} is the batch standard deviation across training processes. The visual policy uses only CLIPScore: rewards are centered within each six-image group and divided by $\max(\sigma_i, 10^{-6})$, where σ_i is the within-prompt standard deviation. These signals are then standardized over all collected images across training processes, using the same 10^{-6} denominator floor, and clipped to $[-5, 5]$. Cached checklists are constructed from each instruction and its training reference explanation, shared across its candidates, and used by the training-time evaluator; Appendix B gives the scoring rules.

Policy updates and evaluation. Each policy makes two optimization passes over the collected rollouts before fresh trajectories are sampled. Visual advantages and old transition log probabilities remain fixed across both visual optimization passes. Both use AdamW with the settings in Table 7. The language policy has KL regularization to its frozen base checkpoint; the visual policy has no KL penalty. Clipped probability ratios and gradient-norm clipping constrain the updates. Evaluation uses 30 denoising steps at the same resolution, with seed 42 and guidance scale 4.5; benchmark-specific protocols are given in Appendix D.1. HPSv3 (Ma et al., 2025) is used in the baseline adaptations and reward ablations. The gradient probes in Appendix F.1 use a distinct reward configuration and update schedule and are interpreted separately from the benchmark comparisons in Table 4.

Table 7: Training hyperparameters for the ThinkGen-based DUETGRPO configuration. Ratio clipping specifies the half-width around 1.

Setting	Language policy	Visual policy
Parameter updates	LoRA	Full DiT
LoRA rank / scaling / configured dropout	64 / 128 / 0.05	–
Samples per conditioning input	$G = 4$	$K = 6$
Optimizer	AdamW	AdamW
Learning rate	10^{-5}	10^{-5}
Adam (β_1, β_2)	$(0.9, 0.999)$	$(0.9, 0.999)$
Adam epsilon / weight decay	10^{-8} / 10^{-4}	10^{-8} / 10^{-4}
Optimization passes per online epoch	2	2
Ratio clipping ϵ_m	0.1	2×10^{-4}
KL coefficient	0.01	0
Maximum gradient norm	1.0	1.0
Optimization microbatch (images)	–	6
Gradient accumulation steps	–	3

D Additional Method and Experimental Details

D.1 Evaluation Protocols

WISE (Niu et al., 2025). WISE (Niu et al., 2025) contains 1,000 prompts spanning six categories of world-knowledge reasoning. We use Qwen3.5-35B-A3B as the judge under the official evaluation protocol (PKU-YuanGroup, 2026), generating one image per prompt. The judge assigns a binary score to each image based on the prompt and its reference explanation. Category scores are the mean per-image scores, and Overall WiScore is their weighted mean: 0.40 for Cultural and 0.12 each for Time, Space, Biology, Physics, and Chemistry. All WISE evaluations performed in this work, including the SFT and GRPO checkpoints, BAGEL experiment, text-reward study, and joint-training controls, use this same judge and scoring protocol.

T2I-CoReBench (Li et al., 2025). T2I-CoReBench (Li et al., 2025) covers four compositional and eight reasoning dimensions. Following its official protocol, we use the official checklist and Gemini 2.5 Flash with temperature zero on the $[0, 100]$ scale. For our SFT and GRPO checkpoints,

we generate four images per prompt with matched seeds, yielding 4,320 images and 54,144 checklist items per model. We average checklist scores per image, images per dimension, and all 12 dimensions for Overall. Composition and Reasoning are the means over their four and eight dimensions, respectively; Overall therefore weights these group means by $4/12$ and $8/12$ before rounding to one decimal place. Reference results use the Gemini 2.5 Flash scores from the official online leaderboard (T2I-CoReBench Team, 2026).

D.2 Joint Training and Reward Comparison Groups

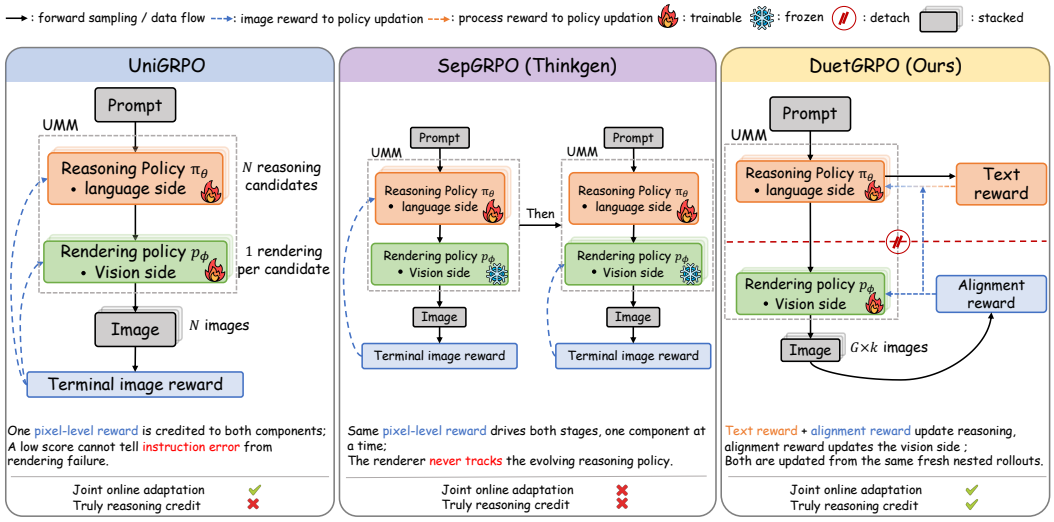


Figure 5: Training approaches for reasoning-driven image generation. ThinkGen alternates updates while freezing one policy. Shared terminal image rewards update both policies from rendered-image scores. DuetGRPO updates both policies in the same online loop: the language policy uses reasoning and image–text alignment rewards, the visual policy uses the same image–text alignment reward, and conditioning is detached to keep policy gradients decoupled.

An image score for y_{ij} depends on both the language policy's generation prompt v_i and the visual policy's rendering trajectory τ_{ij}. Comparing images with a fixed generation prompt measures variation in rendering, but cannot establish whether that generation prompt satisfies the user instruction. Across generation prompts, image scores also reflect differences in what the visual policy can realize. Consequently, an image–text alignment reward computed against the generation prompt alone cannot distinguish an incorrect generation prompt rendered faithfully from a correct generation prompt rendered faithfully.

DuetGRPO uses these comparisons for different purposes. The composite text reward jointly assesses reasoning and whether the generation prompt satisfies the user instruction. The mean image–text alignment reward over images for each generation prompt informs the language policy about the evolving visual policy's ability to follow that generation prompt. The visual-policy advantage compares alignment rewards only among images for the same generation prompt. This does not make image scores a direct measure of reasoning correctness; feedback on reasoning and prompt correctness provides that supervision. Both policies continue to adapt online, while detached conditioning prevents cross-modality gradient coupling.

D.3 Comparison and Control Protocols

Comparison methods. Tables 1 and 2 separate published reference scores from our SFT and GRPO evaluations. "w/ CoT" and "w/ Think" denote explicit reasoning before image synthesis; "DuetGRPO (SFT)" is the supervised initialization, and "DuetGRPO (GRPO)" is the complete joint RL configuration. T2I-CoReBench (Li et al., 2025) reference scores come from the official Gemini 2.5 Flash leaderboard (T2I-CoReBench Team, 2026). WISE (Niu et al., 2025) reference re-

sults follow the cited evaluations and leaderboard (Jiao et al., 2025; PKU-YuanGroup, 2026); in particular, the BAGEL w/ Think and FLUX.1-dev WISE reference scores use the Qwen3.5-35B-A3B leaderboard (PKU-YuanGroup, 2026). All WISE evaluations performed in this work use Qwen3.5-35B-A3B, as specified in Appendix D.1.

The BAGEL experiment applies the complete training recipe to a MoT architecture. The frozen-visual-policy experiment in Table 3(b) uses R^{text} alone; *Freeze vision* in Table 4 instead retains $R^{\text{text}} + \overline{R}^{\text{align}}$. Table 4 follows the common-backbone, common-data setup in Section 4.1. PromptRL, UniGRPO, and Dual-GRPO are ThinkGen adaptations with the implemented reward combinations stated explicitly. Their scores use our common evaluation protocol, and the comparison concerns these implemented configurations.

Baseline reward and update configurations. PromptRL uses HPSv3 for both policies with detached conditioning; UniGRPO uses shared CLIPScore with coupled conditioning gradients; Dual-GRPO uses CLIPScore for language and CLIPScore plus HPSv3 for vision, with detached conditioning. Following their original consistency-reward formulations, UniGRPO and Dual-GRPO compute CLIPScore between each generated image y and the original user instruction x, using $V_{\text{CLIP}}(y, x)$, as specified in Section 4.1. DUETGRPO and its controls use $V_{\text{CLIP}}(y, v)$ for alignment with the generation prompt. The baseline adaptations preserve the original consistency-reward inputs and reward sharing or stage-specific assignment while adapting the scoring models to our comparison protocol.

Our PromptRL adaptation follows the original prompt-retention mechanism (Wang et al., 2026b): images conditioned on original prompts and language-refined prompts enter the same group for reward normalization and advantage estimation. Only refined-prompt samples contribute language-policy gradients, while both types contribute visual-policy gradients.

What each control changes. Relative to DUETGRPO, *Language without text reward* removes only R^{text} from the language reward, retaining $\overline{R}^{\text{align}}$ for language and R^{align} for vision. Alignment still compares each image with its generation prompt v, and language rewards are averaged over the same prompt's renderings before group normalization. *Language without alignment* instead removes only $\overline{R}^{\text{align}}$ from the language reward. Both controls retain joint policy updates, detached conditioning, and language-policy KL regularization under the common training setup. They isolate the contribution of each language-reward component while preserving the visual reward and update configuration. *Freeze vision* holds the visual policy fixed while retaining the full language reward; image sampling and alignment scoring remain necessary to compute that reward. *Freeze language* holds the language policy fixed and updates the visual policy with CLIPScore. *Coupled* retains the complete reward assignment but allows the visual loss to differentiate through language-derived conditioning. These definitions distinguish the current coupled control from the text-reward-only coupled runs in Appendix F.

Figure 3 shows recorded image-reward trajectories over epochs 0–40, and Table 4 reports benchmark generation performance.

E QUALITATIVE COMPARISONS

We provide the complete set of selected paired examples referenced in Section 4.2.1, including the verbatim benchmark prompts, final decoded generation prompts, per-example scores, and same-seed SFT–GRPO image comparisons.

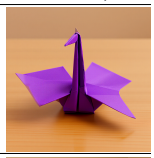
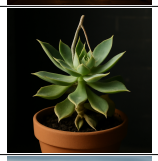
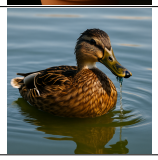
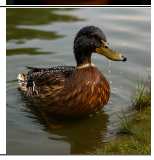
Verbatim benchmark prompt and final decoded generation prompt	SFT	GRPO (text-only reward)
T2I-CoReBench (Li et al., 2025) R-PR-068 (procedural) *Prompt:* Illustrate the final scene after performing all of the following six steps in order: 1. Place a square sheet of purple origami paper flat on a wooden table. 2. Fold the paper diagonally corner to corner and crease sharply, then unfold. 3. Fold along the other diagonal and crease, then unfold to reveal an X-shaped crease pattern. 4. Collapse the paper inward along the creases to form a square base. 5. Continue folding to create the traditional bird base, then pull out the neck, head, and two wings to form a crane. 6. Spread the wings gently so the crane stands upright and centre it on the table. Render the tabletop exactly as it appears once all six steps are complete. **SFT decode (0.600):** Generate an image of a completed origami crane standing upright with its wings spread out, resting in the center of a wooden table. **GRPO decode (1.000):** Generate an image of a completed **purple** origami crane standing upright with its wings spread, centered on a flat wooden table.		
T2I-CoReBench (Li et al., 2025) R-RR-003 (reconstructive) *Prompt:* Observations: A dining table shows the aftermath of a meal. Multiple empty plates with crumbs and sauce stains are stacked at one end. Wine glasses contain only red residue at the bottom. Candles have burned down to small stubs with hardened wax pooled around their bases. A serving platter holds only small portions of leftover food. Chairs are pushed back from the table at various angles. Used napkins are crumpled and scattered across the table surface. Generative Task: Generate an image showing this dining setup at the moment when everything is ready, but the guests have not yet arrived. All objects mentioned in Observations must be reconstructed in the scene, except those that are meant to have disappeared in the reconstructed moment. **SFT decode (0.600):** Generate an image of a dining table set for a meal before guests arrive, featuring **full plates of fresh food**, full wine glasses filled with red wine, **tall lit candles** with fresh wax, a serving platter filled with a large abundant meal, folded napkins placed neatly on the table, and chairs pushed in straight toward the table. **GRPO decode (0.800):** Generate an image of a dining table set for a meal before guests arrive, featuring multiple **clean and empty plates** arranged neatly, multiple wine glasses filled with red wine, multiple **tall unlit candles** with fresh wax, a serving platter filled with a generous, untouched meal, folded napkins placed neatly on the table, and chairs pushed in tightly around the table.		
T2I-CoReBench (Li et al., 2025) R-BR-031 (behavioral) *Prompt:* Generate a photorealistic image depicting the moment after a person's arm has carelessly swept across a kitchen counter. The initial scene was a clean counter with a tight cluster of items: a tall glass full of orange juice, a full, open cardboard carton of milk, and an open glass jar of strawberry jam. Next to them, a slice of plain bread was waiting in an upright toaster. On a separate small plate, positioned several inches away from this main cluster, sat a single ripe banana. The person's sweeping motion struck the cluster of items, sending the glass, the milk carton, and the jam jar tumbling off the counter's edge, while also pushing down the toaster's activation lever. The image should depict the scene after all resulting effects have completely finished. **SFT decode (0.333):** Generate an image of a photorealistic kitchen scene showing the aftermath of a sweeping motion, featuring a **person's arm visible** in the upper frame having just finished sweeping across a counter, where on the counter sits an upright toaster with its activation lever pushed down containing a slice of plain bread that is actively toasting, next to a small plate holding a single undisturbed ripe banana, while on the floor below lie the completely finished effects of the impact, including shattered glass shards, a crushed cardboard milk carton, a puddle of spilled milk, a broken glass jar, a puddle of spilled strawberry jam, and a puddle of spilled orange juice. **GRPO decode (1.000):** Generate an image of a photorealistic kitchen scene showing the aftermath of a spill, featuring a clean kitchen counter with a small plate holding a single ripe banana, an upright toaster with its activation lever pushed down containing a slice of plain bread, and on the floor below the counter, the completely empty, **tipped-over remains** of a tall glass with spilled orange juice, an empty, tipped-over cardboard milk carton with spilled milk, and an empty, tipped-over glass jar with spilled strawberry jam.		
WISE (Niu et al., 2025) 1000 (chemistry) *Prompt:* Potassium permanganate crystal dissolving in water **SFT decode (0.000):** Generate an image of deep purple potassium permanganate crystals dissolving in a clear glass beaker of water, showing the vibrant color spreading through the liquid as the solid crystals break apart and swirl. **GRPO decode (1.000):** Generate an image of deep purple potassium permanganate crystals dissolving in a clear glass beaker of water, showing vibrant magenta plumes spreading through the liquid in a laboratory setting.		
WISE (Niu et al., 2025) 439 (time) *Prompt:* The alleys around Jama Masjid in Delhi when London is reaching noon. **SFT decode (0.000):** Generate an image of the historic stone alleys around Jama Masjid in Delhi **at night**, illuminated by warm streetlights and the soft glow of mosque windows under a dark sky. **GRPO decode (1.000):** Generate an image of the narrow, red sandstone alleys surrounding the intricate carvings of Jama Masjid in Delhi, illuminated by the warm, golden light of **late afternoon at 5:30 PM**, with long shadows stretching across the stone surfaces.		
WISE (Niu et al., 2025) 665 (biology) *Prompt:* A potted succulent placed in a completely dark closet for one week **SFT decode (0.000):** Generate an image of a potted succulent that has become pale, stretched, and etiolated after being in a dark closet for a week. **GRPO decode (1.000):** Generate an image of a potted succulent in a dark closet, showing the plant's **pale, elongated, and drooping stems and leaves** as a result of being in complete darkness for a week.		
WISE (Niu et al., 2025) 874 (physical knowledge) *Prompt:* A duck emerging from a pond after brief submersion **SFT decode (0.000):** Generate an image of a duck emerging from a pond, featuring wet feathers and a dripping beak to indicate it has just come out of the water. **GRPO decode (1.000):** Generate an image of a duck emerging from the edge of a pond, with its head and neck out of the water and its feathers visibly wet with **glistening water droplets**.		

Figure 6: Selected qualitative comparisons of text-only GRPO on ThinkGen. Each row shows the original benchmark prompt, the final decoded generation prompt, and images generated with the same per-case seed. Boldface highlights prompt changes. Scores are per-example WISE (Niu et al., 2025) or T2I-CoReBench (Li et al., 2025) evaluator scores on $[0, 1]$.

F Additional Conditioning-Gradient Diagnostics

We analyze conditioning gradients in archived coupled runs to characterize their contributions to language updates and optimizer-step frequency.

F.1 How Conditioning Gradients Enter Language-Policy Optimization

Scalar feedback and conditioning gradients are different update paths. In loss-minimization notation, let g_T be the language-surrogate gradient without KL, g_{KL} the weighted KL gradient, and $h_\theta(z)$ the differentiable visual conditioning for sampled language output z. With sampled trajectories and scalar advantage weights fixed, the visual loss contributes

$$g_V = \left(\frac{\partial h_\theta(z)}{\partial \theta}\right)^\top \nabla_h \mathcal{L}_{\mathrm{vis}}. \tag{5}$$

For a scheduled language update, at the same model state and on the same sampled data, the gradients supplied before clipping and Adam are

$$g_{\mathrm{detach}} = g_T + g_{\mathrm{KL}}, \qquad g_{\mathrm{coupled}} = g_T + g_{\mathrm{KL}} + g_V. \tag{6}$$

Nonzero g_V contributes an additional visual-loss direction to the *effective optimization gradient*. In DuetGRPO, scalar alignment feedback already enters the language advantage (Eq. (4)); detaching conditioning preserves the language-surrogate and KL gradients together with this feedback and online co-adaptation.

A common-checkpoint directional intervention. We fork the archived coupled CLIPScore run at checkpoint 20, restoring model, optimizer, and random-generator states, and train through epochs 20–39 to checkpoint 40. Both branches use R^{text} alone for language and CLIPScore for vision, with the same update schedule and KL coefficient 0.01. They differ in the visual-gradient intervention, defined using the KL-free reference g_T:

$$\begin{aligned} g_V^{\mathrm{proj}} &= g_V - \frac{\min\{\langle g_T, g_V\rangle, 0\}}{\|g_T\|_2^2}\, g_T, \\ g_V^{\mathrm{norm}} &= \frac{\|g_V^{\mathrm{proj}}\|_2}{\|g_V\|_2}\, g_V. \end{aligned} \tag{7}$$

Projection removes the negatively aligned component; *Norm control* preserves the original direction at the projected norm. Both branches retain coupled conditioning and scheduled language and KL updates; norm matching uses each branch's current gradients.

Nonzero visual gradients reach the language parameters. We record synchronized full-vector statistics over language LoRA parameters for all $2{,}400$ steps per branch. On the 200 scheduled text-update steps, mean $\|g_V\|_2/\|g_T\|_2$ is 2.52% for Projection and 2.60% for Norm control (Figure 7a). The visual gradient is nonzero on every recorded step. The remaining $2{,}200$ steps (91.7%) still update language parameters through visual gradients after the text-update queue is exhausted. Their text gradients are intervention probes only, not additional language-loss updates. Coupling thus adds a gradient component on text-update steps and visual-driven updates between them.

Gradient geometry and local update directions. On scheduled text-update steps, raw inner products are negative in 47% and 45% of steps, but mean absolute cosines are only 0.0081 and 0.0075 (Figure 7b). The mean negative projection relative to the text gradient, $\max\{0, -\langle g_T, g_V\rangle\}/\|g_T\|_2^2$, is 0.0086% and 0.0092%, counting nonconflicting steps as zero. Direct cancellation is therefore small. Actual post-Adam displacements, measured every ten steps, satisfy $\langle g_T, \Delta\theta\rangle < 0$ on all 240 samples per branch (40 text-update and 200 probe-only steps). These measurements characterize local first-order descent directions for the KL-free language surrogate.

Training rewards and the supported conclusion. Across post-intervention rollouts at epochs 21–39, Projection and Norm control respectively obtain mean CLIPScore of 0.99732 and 0.98964 and mean composite text rewards of 0.77596 and 0.77892 (Figure 7c–d). Projection has

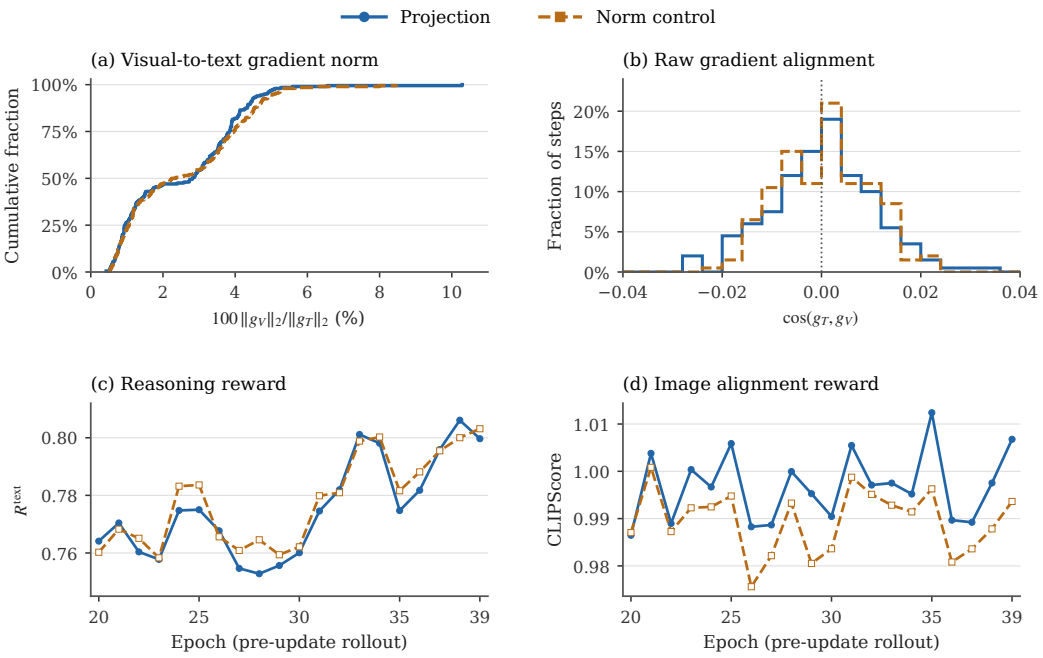


Figure 7: Conditioning gradients from two coupled branches of checkpoint 20. **(a)** Empirical cumulative distribution of the raw visual-to-text gradient norm ratio. **(b)** Raw cosine histogram, normalized by step count. Panels (a–b) use the 200 scheduled text-update steps per branch and a KL-free text gradient. **(c–d)** Unsmoothed rewards over epochs 20–39, measured before each epoch's updates; epoch 20 precedes intervention. Each trajectory represents a single run.

higher CLIPScore in all 19 epochs. Over the final five epochs, the corresponding CLIPScore means are 0.99912 and 0.98843, and the text-reward means are 0.79164 and 0.79368. Each curve tracks one branch from its distinct pre-intervention rollout at epoch 20.

The separate full-reward benchmark comparison favors detached conditioning (0.77 versus 0.75). The diagnostics complement this result by characterizing how conditioning gradients interact with optimizer-step frequency, KL applications, and relative loss scales.