UNIGRPO: UNIFIED POLICY OPTIMIZATION FOR REASONING-DRIVEN VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified models capable of interleaved generation have emerged as a promising paradigm, with the community increasingly converging on autoregressive modeling for text and flow matching for image generation. To advance this direction, we propose a unified reinforcement learning framework tailored for interleaved generation. We validate our approach on its fundamental unit: a single round of reasoning-driven image generation, where the model first expands the user prompt through reasoning, followed by image synthesis. Formulating this multimodal generation process as a Markov Decision Process with sparse terminal rewards, we introduce UniGRPO to jointly optimize text and image generation policies using GRPO. Adopting a minimalist methodology to avoid over-design, we leverage established training recipes for both modalities by seamlessly integrating standard GRPO for reasoning and FlowGRPO for visual synthesis. To ensure scalability to multi-round interleaved generation, we introduce two critical modifications to the original FlowGRPO: (1) eliminating classifier-free guidance to maintain linear, unbranched rollouts, which is essential for scaling to complex scenarios involving multi-turn interactions and multi-condition generation (e.g., editing); and (2) replacing the standard latent KL penalty with an MSE penalty directly on the velocity fields, providing a more robust and direct regularization signal to mitigate reward hacking. Our experiments demonstrate that this unified training recipe significantly enhances image generation quality through reasoning, providing a robust and scalable baseline for future post-training of fully interleaved models.

1 INTRODUCTION

The evolution of generative AI is rapidly progressing toward unified multimodal models (Deng et al., 2025; Wei et al., 2025; Zhao et al., 2025; Xie et al., 2024; Zhou et al., 2024) capable of interleaved generation (Liao et al., 2025). A pivotal advantage of this emerging paradigm is the potential to effectively leverage test-time compute through iterative reasoning—refining prompts, generating images, and reflecting on outputs across multiple rounds to tackle complex image synthesis tasks (Huang et al., 2025b). As the boundaries between modalities blur, the community is increasingly gravitating toward a robust architectural synergy: Autoregressive (AR) (Radford et al., 2019) models for text generation paired with Flow Matching (Lipman et al., 2022; Liu et al., 2022) for visual synthesis (Deng et al., 2025; Xie et al., 2024; Zhou et al., 2024; Liao et al., 2025). This combination harnesses the reasoning capabilities of Large Language Models (LLMs) alongside the high-fidelity generation strengths of Flow-based models.

In this work, we argue that advancing interleaved generation requires a unified Reinforcement Learning (RL) framework that jointly optimizes text and image generation policies. Rather than immediately scaling to long-horizon multi-turn generation, we validate our framework on its fundamental unit: a single round of reasoning-driven image generation. This setting already encompasses both text and image generation, covering the essential components of interleaved generation. In the absence of open-source base models natively capable of full interleaved generation, it serves as a meaningful and principled testbed for validating our unified RL framework.

To this end, we propose UniGRPO, a unified RL framework formulating the entire "Prompt $\rightarrow$ Thinking $\rightarrow$ Image" sequence as a single Markov Decision Process (MDP) (Puterman, 1990). Adopting a minimalist methodology to avoid over-design, we integrate established training recipes

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.

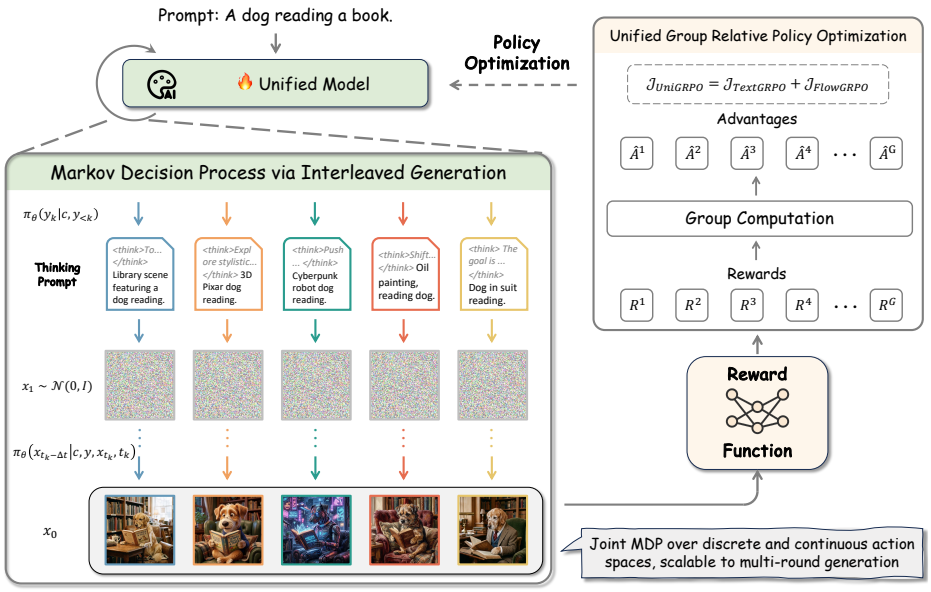

Figure 1: **Overview of UniGRPO.** By formulating interleaved generation as a joint MDP, this illustration demonstrates how UniGRPO jointly optimizes discrete language actions (y_k) in the LLM's next-token prediction, and continuous visual actions ($x_{t_k-\Delta t}$) in flow matching. Both policies are updated using group-relative advantages derived from sparse terminal rewards.

for both modalities: standard GRPO (Shao et al., 2024) for the reasoning component and Flow-GRPO (Liu et al., 2025) for visual synthesis. Under sparse terminal rewards, UniGRPO jointly optimizes both text and image generation policies, encouraging the model to produce more informative reasoning texts while simultaneously improving the visual synthesis process itself.

Crucially, our design choices are driven by the goal of scalability to future multi-round and multi-condition scenarios (e.g., complex editing tasks). We introduce two critical modifications to the standard Flow Matching RL training recipe within our framework. First, we eliminate Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) during training. While CFG is a standard inference technique, its removal ensures that the generation process remains a linear, unbranched rollout, which is essential for scaling to complex scenarios involving multi-turn interactions and multi-condition generation. Second, we replace the standard latent KL penalty with an MSE penalty directly on the velocity fields. This provides a more robust and direct regularization signal that effectively mitigates reward hacking, ensuring the optimization remains well-grounded. Our contributions can be summarized as follows:

- Unified RL Framework for Reasoning-Driven Image Generation: We propose UniGRPO, a minimalist framework that formulates the Prompt → Thinking → Image sequence as a single MDP, jointly optimizing AR text and flow-matching image policies. We validate this framework on the fundamental unit of interleaved generation, demonstrating that jointly optimizing reasoning and visual synthesis improves image generation quality.
- Scalable Flow Matching RL Adaptations: We introduce two critical modifications to FlowGRPO: eliminating CFG to ensure unbranched rollouts, and replacing the standard latent KL penalty with an MSE penalty directly on the velocity fields for more robust reward hacking mitigation. Together, these adaptations are essential for scaling to multi-turn and multi-condition scenarios.
- We demonstrate that our unified training recipe effectively optimizes the model under sparse terminal rewards, establishing a scalable baseline for future post-training of fully interleaved models.

2 Related Work

2.1 RL for LLMs

Recent LLM advancements rely on Reinforcement Learning (RL) for alignment and reasoning. While PPO (Schulman et al., 2017) is a standard approach, the highly efficient Group Relative

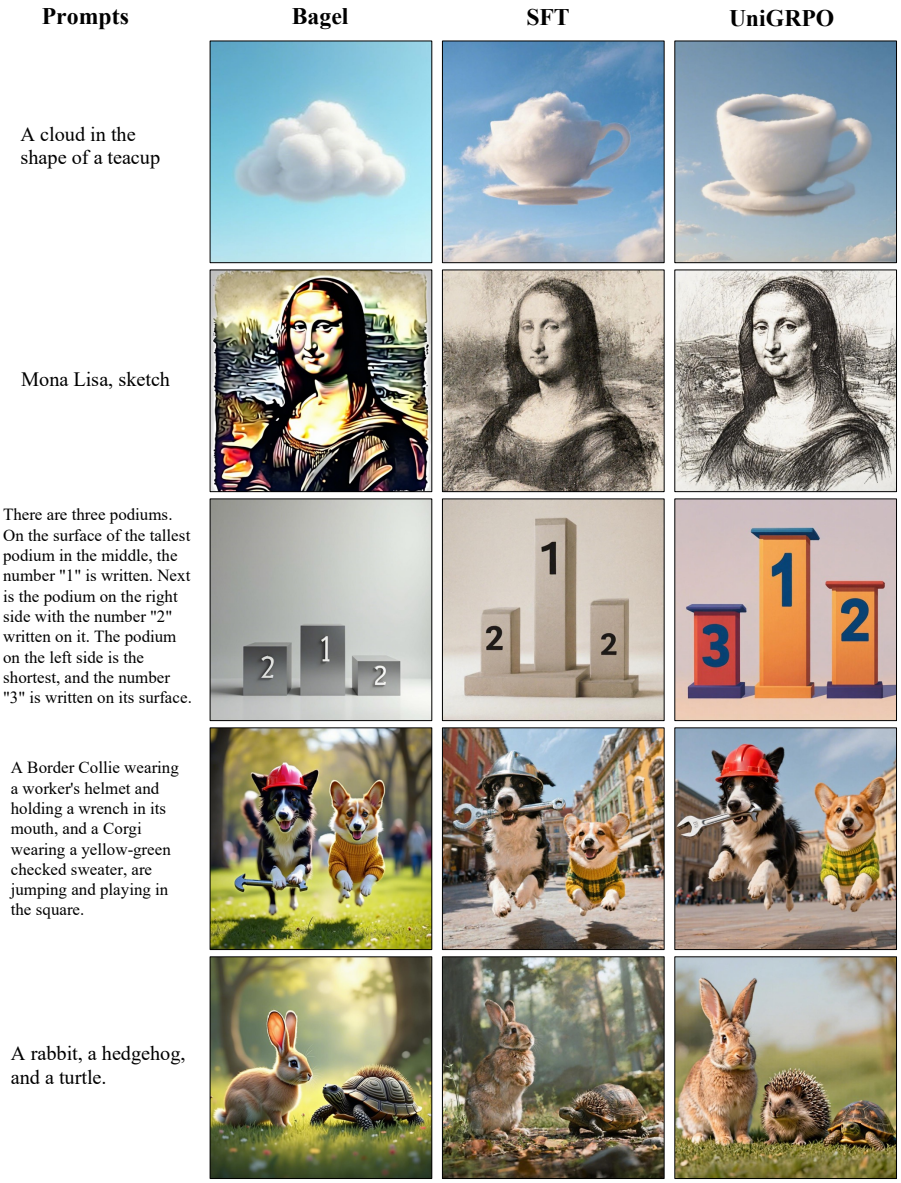

Figure 2: T2I qualitative comparison.

Policy Optimization (GRPO) (Shao et al., 2024) eliminates the value model by using group-relative baselines. This efficiency drives reasoning-intensive models using Chain-of-Thought (CoT) (Wei et al., 2022), such as DeepSeek-R1. Our work adapts GRPO to efficiently optimize the intermediate "thinking" tokens prior to visual synthesis.

2.2 RL for Diffusion and Flow Matching Models

Aligning text-to-image (T2I) models with human intent has been extensively explored via reward-gradient optimization (Prabhudesai et al., 2023; Clark et al., 2023; Xu et al., 2024), reward-weighted regression (Peng et al., 2019; Lee et al., 2023; Fan et al., 2025), Direct Preference Optimization (DPO) (Rafailov et al., 2024; Wallace et al., 2024; Liang et al., 2024), and PPO-style policy gradients (Black et al., 2023; Fan et al., 2024). Applying policy gradients to the deterministic ODEs of flow matching models requires a stochastic formulation: FlowGRPO (Liu et al., 2025) and DanceGRPO (Xue et al., 2025b) reformulate generation as an SDE, and follow-up works (Huang et al., 2025a; Zheng et al., 2025; Deng et al., 2026; Yang et al., 2025; He et al., 2025) further improve training stability, reward design, and sample efficiency. We extend this line of work to jointly optimize language reasoning and visual synthesis.

2.3 Unified Multimodal Understanding and Generation Models

Multimodal understanding and image generation have long evolved independently, dominated by autoregressive and diffusion models, respectively; recent work seeks to unify both within a single framework. One line of research quantizes visual signals so that image and text tokens share an autoregressive space, as in Chameleon (Team, 2024), Emu3 (Wang et al., 2024), and VILA-U (Wu et al., 2024). Another combines autoregressive and diffusion objectives: Show-o (Xie et al., 2024) and Transfusion (Zhou et al., 2024) train a single transformer with mixed next-token prediction and diffusion losses, while Bagel (Deng et al., 2025) and Mogao (Liao et al., 2025) scale this hybrid paradigm with large-scale interleaved data, demonstrating strong complex reasoning and interleaved text-image generation. As surveyed by Zhao et al. (2025), key challenges remain in tokenization, cross-modal attention design, and training data construction.

2.4 Concurrent Work

Several concurrent studies also apply RL to unified or joint multimodal generation. R3 (Ye et al., 2026) uses a generate-understand-regenerate loop but is validated on benchmark-specific prompts rather than general-purpose training. DualGRPO (Kou et al., 2026) jointly optimizes a separate LLM and diffusion backbone via tree-structured rollouts, a design incompatible with true interleaved generation. PromptRL (Wang et al., 2026) likewise trains disjoint language and flow models in a joint RL loop, but on limited training data. SepGRPO (Jiao et al., 2025), also built on BAGEL, alternates RL between the MLLM and DiT modules rather than optimizing them jointly end-to-end. In contrast, our method trains a single unified model end-to-end with general-purpose prompts at 1024 resolution, using a scalable algorithm built upon an improved FlowGRPO, and is compared against a wide range of diffusion RL baselines.

3 Preliminary

In this section, we establish the theoretical foundations for optimizing generative policies using Unified Group Relative Policy Optimization (UniGRPO), covering both discrete text generation and continuous flow-based visual generation.

3.1 Text GRPO

For the autoregressive text component, we adopt the standard GRPO (Shao et al., 2024) formulation. Given a prompt c, the policy π_θ generates a group of G outputs $\{y_i\}_{i=1}^{G}$. The objective maximizes the expected reward while constraining policy updates via importance-ratio clipping.

The advantage for the i-th sample is computed relatively within the group:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}. \tag{1}$$

The objective function is defined as:

$$\mathcal{J}_{\text{Text}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{k=1}^{|y_i|}\left(\min\left(r_{i,k}\,\hat{A}_i,\ \text{clip}(r_{i,k}, 1{-}\epsilon, 1{+}\epsilon)\,\hat{A}_i\right) - \beta_{\text{txt}}\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right), \tag{2}$$

where $r_{i,k} = \frac{\pi_\theta(y_{i,k}|y_{i,<k})}{\pi_{\theta_{\text{old}}}(y_{i,k}|y_{i,<k})}$ denotes the importance ratio at step k.

3.2 Flow GRPO

For the visual component, we utilize FlowGRPO (Liu et al., 2025), which adapts reinforcement learning to flow matching models by converting the deterministic Ordinary Differential Equation (ODE) into a Stochastic Differential Equation (SDE) to enable exploration.

SDE Sampling. To introduce the necessary stochasticity for RL exploration, the sampling process is formulated as:

$$\Delta x_{t_k} = \left[v_\theta(x_{t_k}, t_k) + \frac{\sigma_{t_k}^2}{2\, t_k} \big(x_{t_k} + (1 - t_k)\, v_\theta(x_{t_k}, t_k) \big) \right] \Delta t + \sigma_{t_k} \sqrt{\Delta t}\, \epsilon, \tag{3}$$

where σ_{t_k} controls the noise level and $\epsilon \sim \mathcal{N}(0, I)$. For training efficiency, we adopt the FlowGRPO-Fast variant (Liu et al., 2025), which employs a hybrid sampling strategy. Specifically, denoising steps within a continuous time window are performed via SDE and optimized with gradient tracking, while the remaining steps follow standard ODE sampling without gradient computation. This significantly reduces computational cost while preserving optimization effectiveness.

Mitigating Reward Hacking via RatioNorm. Standard importance-ratio clipping often fails in diffusion models because the distribution of importance ratios

$$r_{t_k}(\theta) = \frac{p_\theta(x_{t_k - \Delta t} \mid c, y, x_{t_k}, t_k)}{p_{\theta_{\text{old}}}(x_{t_k - \Delta t} \mid c, y, x_{t_k}, t_k)}, \tag{4}$$

is systematically left-shifted (mean < 1) and exhibits inconsistent variance across timesteps (Wang et al., 2025). This prevents the clipping mechanism from constraining overconfident positive updates, leading to severe reward hacking. To address this, we adopt the Ratio Normalization (RatioNorm) proposed in GRPO-Guard (Wang et al., 2025). It standardizes the log-importance ratio to center its distribution around zero, restoring the effectiveness of clipping:

$$\log \tilde{r}_{t_k}(\theta) = \sigma_{t_k} \sqrt{\Delta t} \left(\log r_{t_k}(\theta) + \frac{\|\Delta \mu_\theta(x_{t_k}, t_k)\|^2}{2\, \sigma_{t_k}^2\, \Delta t} \right), \tag{5}$$

where $\Delta \mu_\theta(x_{t_k}, t_k) \triangleq \mu_{\theta_{\text{old}}}(x_{t_k}, t_k) - \mu_\theta(x_{t_k}, t_k)$ is the mean drift between the two policies.

Combining the hybrid SDE sampling strategy with the RatioNorm mechanism, the final FlowGRPO objective is computed exclusively over the SDE timestep subset $\mathcal{T}_{\text{SDE}}$:

$$\mathcal{J}_{\text{Flow}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\mathcal{T}_{\text{SDE}}|} \sum_{t_k \in \mathcal{T}_{\text{SDE}}} \left(\min \left(\tilde{r}_{i,t_k}\, \hat{A}_i,\ \text{clip}(\tilde{r}_{i,t_k}, 1-\epsilon, 1+\epsilon)\, \hat{A}_i \right) - \beta_{\text{img}}\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right), \tag{6}$$

where $|\mathcal{T}_{\text{SDE}}|$ denotes the number of denoising steps within the continuous SDE window.

4 METHOD

Building upon these foundations, we propose **UniGRPO**, a unified framework that jointly optimizes multimodal generation policies within a single reinforcement learning loop.

4.1 MULTIMODAL GENERATION AS A MARKOV DECISION PROCESS

We formulate interleaved generation as a sequential MDP $(\mathcal{S}, \mathcal{A}, P, R)$, where each MDP step k is a single token prediction in the text phase or a single denoising step in the image phase.

- **State Space $\mathcal{S}$:** The state evolves through two phases. In the *text phase*, $s_k^{\text{txt}} = (c,\ y_{<k})$ comprises the input prompt c and all previously generated reasoning tokens $y_{<k}$. In the *image phase*, $s_k^{\text{img}} = (c,\ y,\ x_{t_k},\ t_k)$ includes the prompt, the completed reasoning trace y, the noisy image latent x_{t_k}, and the current flow time t_k.
- **Action Space $\mathcal{A}$:** In the text phase, $a_k^{\text{txt}} \in \mathcal{V}$ is a single token drawn from the vocabulary. In the image phase, $a_k^{\text{img}} = x_{t_k - \Delta t} \in \mathbb{R}^d$ is the denoised latent at the next flow step.
- **Transition P:** Both phases are deterministic given the action: the text transition appends a_k^{txt} to the sequence, and the image transition moves the latent from x_{t_k} to $x_{t_k - \Delta t}$.
- **Reward R:** A sparse terminal reward $R(x_0, c)$ is assigned only after the image latent has been fully denoised to x_0; all intermediate steps receive zero reward.

4.2 UniGRPO Framework

Given a unified model π_θ that performs interleaved generation, UniGRPO models the entire generation process as a MDP and optimizes it through group relative policy optimization. Specifically, for a given prompt c, we first sample G reasoning chains $\{y_i\}_{i=1}^G$ via $\pi_\theta(a_k^{\text{txt}} \mid s_k^{\text{txt}})$. Each reasoning chain then conditions the same model to generate a corresponding image trajectory $\{x_i\}_{i=1}^G$ via $\pi_\theta(a_k^{\text{img}} \mid s_k^{\text{img}})$ with a hybrid SDE-ODE integrator. We compute group-relative advantages $\hat{A}_i$ based on the terminal rewards of the completed multimodal trajectories. These advantages are used to update π_θ through a unified objective:

$$\mathcal{J} = \mathcal{J}_{\text{Text}} + \lambda \, \mathcal{J}_{\text{Flow}}, \tag{7}$$

where λ is a hyperparameter controlling the relative weight of the image generation objective. To equally balance the reasoning and synthesis tasks, we simply set $\lambda = 1$ across all our experiments. To ensure scalability to multi-round interleaved generation, we introduce two critical modifications to the training recipe.

Eliminating Classifier-Free Guidance. Standard flow matching inference typically relies on CFG to enhance prompt adherence, requiring two model evaluations per step (conditional and unconditional). Crucially, this computational burden scales with the number of conditions; for multi-condition generation such as image editing, CFG demands at least three evaluations per step. Furthermore, this complexity compounds in multi-round interleaved generation, where the system must continuously manage and branch multiple conditional contexts across alternating text and image phases. In an RL setting, this multiplication of function evaluations and context branches drastically inflates computational and memory costs, while creating a branched computation graph that severely complicates gradient estimation. We therefore train UniGRPO entirely without CFG, enforcing a linear, unbranched rollout. While removing CFG typically degrades prompt adherence, our framework compensates for this during training. By explicitly maximizing the expected reward—which evaluates text-image alignment and visual quality—we internalize the alignment capabilities directly into the policy weights. This establishes a highly efficient pipeline that naturally scales to complex multi-condition, multi-round interaction generation.

Velocity-Based Regularization. Preventing reward hacking is a primary challenge in RL for visual generation. In the above SDE formulation, the step-wise transition probabilities are Gaussian, meaning the exact local KL divergence in the latent space can be analytically computed. Specifically, this exact KL evaluates to the squared difference in predicted velocities, weighted by the inverse noise variance ($1/\sigma_{t_k}^2$). However, this inherent weighting applies an uneven penalty across the generative trajectory. For instance, at timesteps with high noise variance, the KL penalty becomes excessively small. This inconsistency creates temporal vulnerabilities that the RL optimizer can easily exploit. To achieve a more robust and consistent constraint, we drop this timestep-dependent weighting and apply a Mean Squared Error (MSE) penalty directly on the unweighted velocity fields:

$$\mathcal{L}_{\text{MSE}}(\theta) = \|v_\theta(x_{t_k}, t_k, y) - v_{\text{ref}}(x_{t_k}, t_k, y)\|^2. \tag{8}$$

This unweighted formulation explicitly forces the RL-tuned vector field to remain close to the pretrained reference model uniformly across all noise levels. Empirically, we find that this uniform regularization leaves fewer loopholes for policy exploitation, proving significantly more effective at mitigating reward hacking while safely preserving the base model's generative priors.

5 Experiments

This section presents the empirical validation of the proposed UniGRPO framework. We begin by outlining the experimental setup—including the pretrained model, reward formulation, baselines, and evaluation protocols. Detailed hyperparameter settings are deferred to Appendix 3. Following this, we compare UniGRPO against strong baselines and conclude with ablation studies to evaluate critical design choices.

5.1 Experimental Settings

The Pretrained Model. As a preliminary exploration into reinforcement learning for interleaved generation, we require a backbone capable of handling mixed-modal outputs. We adopt Bagel (Deng

Table 1: **Main results on TA and GenEval.** All RL methods start from the Bagel checkpoint after SFT. "Thinking" denotes whether the method explicitly generates intermediate reasoning tokens. − indicates training collapse.

Model / Method	Thinking	TA Score	GenEval
Bagel	×	0.6810	0.78
Bagel	✓	0.7132	0.79
SFT	×	0.7486	0.83
SFT	✓	0.7769	0.82
ReFL	×	0.7786	0.85
ReFL	✓	0.8120	0.84
FPO	×	0.7893	0.87
FPO	✓	0.8159	0.85
FlowGRPO	×	0.8112	0.88
FlowGRPO	✓	0.8208	0.86
TextGRPO	✓	0.8078	0.88
ReFL (w/ Thinking)	✓	0.7804	0.83
ReFL (w/ Thinking) + TextGRPO	✓	0.7987	0.87
UniFPO	✓	−	−
UniGRPO (Ours)	✓	**0.8381**	**0.90**

et al., 2025), a model architecture with inherent interleaved generation potential. However, we observed that the vanilla Bagel exhibits limited instruction-following capabilities and suboptimal image generation quality. To establish a strong baseline, we performed Supervised Fine-Tuning (SFT) on Bagel using a curated internal dataset. This process significantly boosted performance (see Table 1). Unless otherwise stated, all subsequent baselines and experiments utilize this finetuned Bagel as the starting checkpoint.

Reward Model. A key advantage of the GRPO algorithm is its flexibility; it does not require differentiable reward functions, allowing the integration of black-box verifiers or VLM-based feedback. However, to ensure a fair comparison with gradient-based baselines like ReFL (Xu et al., 2024) (which necessitates differentiable rewards), we utilize a differentiable reward formulation for the main experiments. Specifically, we employ the exact same reward model as utilized in RewardDance (Wu et al., 2025). This model is fine-tuned based on InternVL (Chen et al., 2024) using collected user preference data, explicitly designed to measure the consistency between generated images and user prompts. It is important to note that while ReFL is restricted to such differentiable objectives, UniGRPO is compatible with a broader range of verifier-based rewards.

Baselines. **ReFL** directly fine-tunes diffusion models by viewing reward model scores as human preference losses and back-propagating gradients to a randomly-picked late timestep t. **ReFL w/ Thinking** generates thinking prompts during training and optimizing only the image generation part using the ReFL objective. **ReFL + TextGRPO** follows a two-stage paradigm: initializing from the trained ReFL w/ Thinking checkpoint and subsequently optimizing the text generation module using TextGRPO. **FPO / AWR** (McAllister et al., 2025; Xue et al., 2025a) serves as an alternative to FlowGRPO. Unlike FlowGRPO which introduces SDE perturbations for exploration, FPO utilizes the forward process to obtain x_t and uses the Evidence Lower Bound (ELBO) of the denoising process as a surrogate for $\log p_\theta(x_0|c)$ to compute importance sampling weights. **UniFPO** denotes a unified framework analogous to UniGRPO, where the text component is optimized via TextGRPO and the image synthesis component is trained using the FPO objective.

Evaluation Metrics. We evaluate generation quality and prompt alignment on two benchmarks:

- Text Alignment (TA) Benchmark: Our internal evaluation set consisting of 150 diverse prompts. For each prompt, we generate 4 images. Evaluation is performed by a VLM, which assesses the outputs against multiple specific exam points defined for each prompt. Each exam point receives a binary score (1 for correct, 0 for incorrect), and the score for a single image is calculated as the average score across all its associated exam points. The final reported metric is the overall average score across all evaluated images. We refer to RewardDance (Wu et al., 2025) for further details on this scoring mechanism.

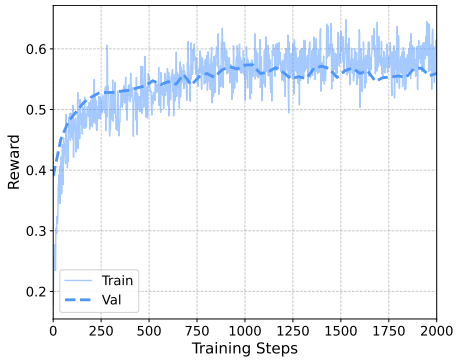

Figure 3: Training and Validation reward curves of UniGRPO on the Finetuned Bagel base model at a resolution of 1024. The x-axis represents the gradient update steps.

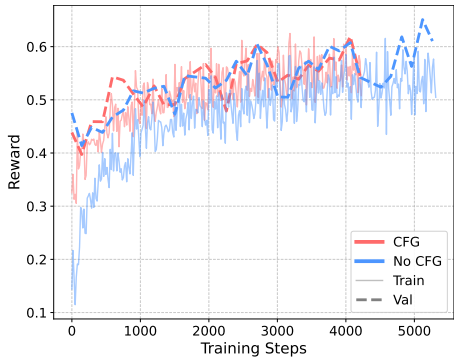

Figure 4: **Ablation Study on CFG.** Training without CFG matches or outperforms training with CFG (CFG is applied at evaluation for all settings). Base model: original Bagel at 512 resolution.

- GenEval (Ghosh et al., 2023): A standard benchmark assessing compositional T2I capabilities, including object counting, spatial relations, and attribute binding.

5.2 Main Results

We begin by analyzing the learning dynamics of UniGRPO, presenting the training and validation reward curves in Figure 3 alongside qualitative generation examples in Figure 2. Next, we benchmark our framework against several established baselines: ReFL, FPO, FlowGRPO, TextGRPO, and hybrid approaches. The quantitative comparisons are summarized in Table 1.

Benchmark Performance. The results in Table 1 indicate that SFT significantly improves the base capabilities of Bagel. Among the RL methods, UniGRPO achieves state-of-the-art performance, scoring 0.8381 on TA and 0.90 on GenEval. Notably, UniFPO failed to converge, underscoring the stability advantages of our GRPO-based formulation. The comparison between UniGRPO, FlowGRPO, and TextGRPO confirms that jointly optimizing both the reasoning and synthesis policies yields gains superior to optimizing either component in isolation. Furthermore, we observe that enabling the explicit reasoning chain ("Thinking") on the Bagel model does not consistently improve GenEval scores. As noted by the Bagel authors, their reasoning module is primarily trained for knowledge-based reasoning and may not be ideally suited for short prompt rewriting tasks found in benchmarks like GenEval[1]. However, our UniGRPO framework successfully leverages the reasoning chain to achieve SOTA performance. Detailed GenEval metrics are in Appendix A.1.

Qualitative Analysis. As illustrated in Figure 2, the original Bagel tends to generate images with oversaturated colors and noticeable synthetic artifacts. While SFT helps mitigate these synthetic artifacts, it compromises image sharpness, resulting in noticeable blurriness upon close inspection. Overcoming this limitation, our proposed UniGRPO significantly enhances both aesthetic quality and text-image alignment, yielding photorealistic, finely detailed images that faithfully reflect complex user prompts. Beyond visual quality, we analyze the models' internal reasoning processes in Appendix A.2 (Figures 6–8). Note that during SFT, we standardized Bagel's original `<think>` format to align with prevalent LLM conventions. While the base Bagel and SFT models generate detailed reasoning texts, these traces can sometimes lose focus or become disconnected from the core visual generation task. In contrast, UniGRPO optimizes the reasoning phase to be highly purposeful and task-oriented. By explicitly aligning the thought process with the final visual reward, UniGRPO produces reasoning traces that tightly couple with and effectively guide image synthesis.

5.3 Ablation Study

We conduct ablation studies to validate our specific architectural and training decisions, including the removal of CFG and the choice of KL regularization.

[1] https://github.com/ByteDance-Seed/Bagel/issues/109#issuecomment-2934226809

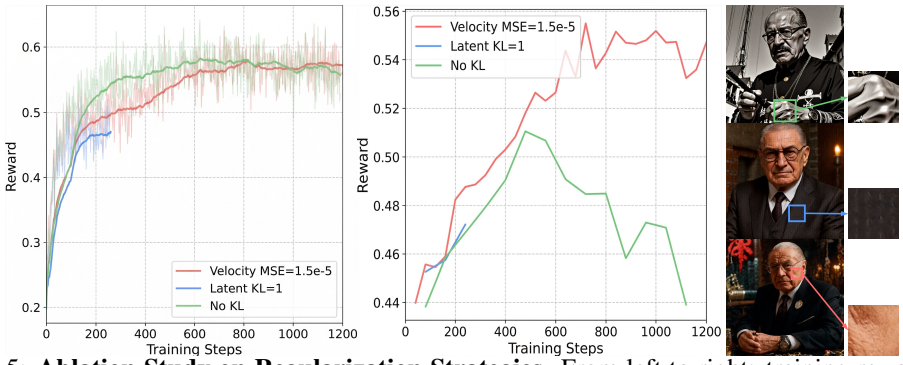

Figure 5: **Ablation Study on Regularization Strategies.** From left to right: training reward, validation reward, and images generated under three different regularization strategies. Without regularization, the validation reward drops after an initial increase, leading to unnatural, oversaturated textures in the generated images. For KL divergence on the latents, the significant drop in training reward indicates that a sufficiently large KL coefficient has been used, yet grid-like artifacts still emerge as early as step 250, prompting us to terminate this run early. In contrast, directly applying MSE regularization on the velocity field ensures stable training dynamics and produces high-fidelity images with realistic textures.

Impact of CFG-Free Training. We compared training UniGRPO with and without CFG. As shown in Figure 4, although CFG during training yields images with higher rewards, removing CFG results in comparable or better final performance when evaluated with CFG. This confirms that CFG is unnecessary for RL-based alignment, rendering the computationally expensive branched CFG rollouts unnecessary during training. Since this ablation uses the original Bagel at 512 resolution, its curves are not directly comparable to Figure 3.

Regularization Strategies. Preventing reward hacking is critical in RL. We compared three strategies: (1) No KL, (2) Latent KL (standard practice), and (3) Velocity MSE (Ours). As shown in Figure 5, removing KL leads to reward hacking where metrics are high but quality degrades. Velocity MSE achieves the best balance, constraining the vector field to the reference model while maintaining strong generation performance.

6 Conclusion and Future Work

In this work, we presented UniGRPO, a unified reinforcement learning framework designed to align interleaved text-and-image generation models. By formulating the multimodal generation process as an MDP, we successfully integrated autoregressive reasoning and flow-based visual synthesis into a single optimization loop. Our minimalist approach establishes a scalable training recipe by eliminating CFG to enforce linear rollouts and employing velocity-based regularization to mitigate reward hacking. Empirically, we demonstrated that UniGRPO effectively enhances image generation quality through chain-of-thought reasoning, providing a scalable baseline for future post-training of fully interleaved models. We identify two promising directions to extend this framework:

Scaling to Multi-Round Interleaved Generation. While this work validates UniGRPO on single-round generation (Prompt $\rightarrow$ Reasoning $\rightarrow$ Image), the unified MDP formulation naturally extends to multi-turn interactions. Future work will apply UniGRPO to complex scenarios such as interactive image editing, visual storytelling, and multi-turn dialogue. In these settings, the model must maintain context consistency across long horizons, where our CFG-free, linear rollout strategy is expected to offer significant computational advantages and training stability.

Multimodal Process Reward Modeling. Currently, UniGRPO relies on sparse terminal rewards evaluated solely on the final generated image. This can lead to inefficient credit assignment, where the model may generate high-quality images despite flawed reasoning. A critical future direction is to introduce *Multimodal Process Reward Models* (PRMs) that provide dense feedback on the intermediate reasoning steps. By verifying whether the generated "thoughts" are logically sound and aligned with the visual intent before the image is generated, we can further improve the sample efficiency of RL training and ensure better interpretability of the model's decision-making process.

AI USE STATEMENT

In this work, we have not used generative AI tools for research ideation, method design, writing code, running experiments, analyzing results, or producing figures and tables; these tasks were carried out entirely by the authors. We used generative AI tools only for LaTeX typesetting (adapting the manuscript to the ICLR 2027 template) and minor copy-editing to meet the page limit (e.g., condensing sentences in the related work section and figure captions), without adding new content or claims. We have reviewed all AI-assisted edits. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

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.

Haoyou Deng, Keyu Yan, Chaojie Mao, Xiang Wang, Yu Liu, Changxin Gao, and Nong Sang. Densegrpo: From sparse to dense reward for flow matching model alignment, 2026. URL https://arxiv.org/abs/2601.20218.

Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with wasserstein regularization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Haoran He, Yuxiao Ye, Jie Liu, Jiajun Liang, Zhiyong Wang, Ziyang Yuan, Xintao Wang, Hangyu Mao, Pengfei Wan, and Ling Pan. Gardo: Reinforcing diffusion models without reward hacking, 2025. URL https://arxiv.org/abs/2512.24138.

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

Tao Huang, Zhao Zhong, Yutao Cui, Chun Fan, Junzhe Li, Yinping Ma, and Miles Yang. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2025a. URL https://arxiv.org/abs/2507.21802.

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

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.

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.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint 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. *arXiv preprint arXiv:2505.05470*, 2025.

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

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. *arXiv preprint arXiv:2507.21053*, 2025.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

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

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

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, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

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.

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*, 2026.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 2025.

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.

Hongyang Wei, Baixin Xu, Hongbo Liu, Size Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 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.

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*, 2025.

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.

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

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.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025a.

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*, 2025b.

Jian Yang, Bo Zhang, Siming Fu, Dacheng Yin, Fengyun Rao, Yuke Zhao, Xiaoxuan He, and Wanli Li. Tempflow-grpo: When timing matters for grpo in flow models, 2025. URL https://arxiv.org/abs/2508.04324.

Sen Ye, Mengde Xu, Shuyang Gu, Di He, Liwei Wang, and Han Hu. Understanding vs. generation: Navigating optimization dilemma in multimodal models. *arXiv preprint arXiv:2602.15772*, 2026.

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.

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, 2025. URL https://arxiv.org/abs/2509.16117.

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

A Extended Experimental Results

A.1 Detailed GenEval Results

Table 2 provides the comprehensive category-level breakdown of the GenEval benchmark. We present the fine-grained metrics across all six sub-categories (Single Object, Two Objects, Counting, Colors, Position, and Attribute Binding) for all evaluated models and baselines.

Table 2: Main quantitative results on the GenEval benchmark. All RL methods use the Bagel checkpoint after SFT. "Thinking" denotes whether the method explicitly generates intermediate reasoning tokens. × indicates training collapse.

Model / Method	Thinking	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Bagel	×	0.78	0.98	0.96	0.78	0.84	0.52	0.58
Bagel	✓	0.79	0.99	0.92	0.77	0.88	0.56	0.62
SFT	×	0.83	0.99	0.95	0.83	0.89	0.58	0.75
SFT	✓	0.82	0.98	0.93	0.63	0.91	0.68	0.79
ReFL	×	0.85	1.00	0.97	0.86	0.92	0.57	0.81
ReFL	✓	0.84	0.99	0.96	0.63	0.94	0.70	0.82
FPO	×	0.87	0.99	0.99	0.90	0.93	0.59	0.86
FPO	✓	0.85	0.99	0.97	0.69	0.91	0.69	0.81
FlowGRPO	×	0.88	0.99	0.98	0.93	0.94	0.60	0.86
FlowGRPO	✓	0.86	0.99	0.96	0.76	0.90	0.71	0.84
TextGRPO	✓	0.88	0.99	0.96	0.87	0.91	0.76	0.84
ReFL (w/ Thinking)	✓	0.83	0.99	0.94	0.64	0.92	0.70	0.81
ReFL (w/ Thinking) + TextGRPO	✓	0.87	0.98	0.97	0.84	0.91	0.75	0.80
UniFPO	✓	×	×	×	×	×	×	×
UniGRPO (Ours)	✓	**0.90**	0.99	0.99	0.91	0.91	0.73	0.86

A.2 Qualitative Results Across Training Stages

We present qualitative comparisons of reasoning traces and generated images across different training stages. Figure 6 shows the reasoning trace and generated image of the baseline Bagel model. Figure 7 illustrates the result after SFT, during which we standardized the original `<think>` format to match prevalent LLM conventions. Although the SFT model produces extensive reasoning steps, the translation from textual thoughts to fine-grained visual details remains suboptimal. Figure 8 demonstrates that the subsequent RL-based training stage effectively addresses this gap, yielding more coherent reasoning traces and higher-quality image generation.

B Implementation Details and Hyperparameters

Table 3 details the comprehensive hyperparameter settings used for the joint training of our Mixture of Experts (MoT) architecture. This includes the specific optimization configurations for both the text reasoning expert (TextGRPO) and the image denoising expert (FlowGRPO), alongside the joint objective weight λ.

Prompts	Thinking Texts	Images
Six cups are arranged in two rows, with three in each row, viewed from above.	<think> The model should generate an image showing the cups from a top-down perspective, with the arrangement clearly visible as two rows of three cups each. The expanded prompt follows: Six ceramic cups arranged in two neat rows of three, viewed from a top-down perspective, showcasing their symmetrical arrangement and clean, minimalist design. The cups are evenly spaced, with a neutral background that emphasizes their form and arrangement. The lighting is soft and diffused, creating subtle shadows and a calm, balanced atmosphere. The image is ultra-realistic, with fine details visible in the texture of the ceramic and the smooth surfaces of the cups. </think>	
A boy is holding a piece of wood, and a girl is drawing on it with a pen.	<think> The model should generate an image where the girl is actively drawing on the wood held by the boy, showing interaction between them and the object. Here's the full, detailed prompt: A young boy stands holding a piece of wood horizontally, while a girl, seated nearby, actively draws on the wood with a pen. The scene is set in a bright, cozy workshop filled with natural light streaming through large windows, casting soft shadows. The wood has visible grain patterns, and the girl's hand is steady as she sketches intricate designs. The boy appears curious and supportive, watching her work with a gentle smile. The atmosphere is warm and creative, with subtle hints of wood shavings and tools scattered around the workspace. The image is highly detailed, ultra-realistic, with vibrant colors and a sense of realism, capturing the interaction between the two characters and the wooden surface. </think>	
A large cat climbing frame has three platforms. On the three platforms, there is a cat, a dog and a mouse lying respectively.	<think> The model should generate an image of a cat climbing frame with three platforms, each occupied by a cat, a dog, and a mouse, demonstrating a playful interaction among different animals. Here's the finished detailed prompt: A large, detailed cat climbing frame made of sturdy wood and metal, featuring three distinct platforms of varying heights. On the lowest platform, a relaxed cat lies sprawled out, its fur soft and fluffy, embodying comfort. On the middle platform, a playful dog sits upright, its ears perked and tail wagging slightly, exuding energy and curiosity. On the highest platform, a tiny mouse lies curled up, its tiny whiskers twitching softly, appearing calm and secure. The scene is set in a bright, airy environment with natural light streaming through large windows, casting soft shadows and creating a warm, inviting atmosphere. The image is ultra-realistic with fine details, showcasing textures like the rough wood grain, the smooth fur of the animals, and the intricate design of the climbing frame. </think>	
On the grassland, a tiger is chasing two sheep, one big and one small.	<think> The model should generate an image where the tiger is depicted as a predator and the sheep are prey, emphasizing the chase dynamic. Here's the expanded prompt: A dynamic and vivid scene set on a vast, sunlit grassland under a clear blue sky, featuring a powerful tiger in mid-chase, its muscles tense and focused, as it pursues two sheep—one large and one small—scrambling away in panic. The tiger is depicted as a formidable predator with realistic fur texture and lifelike detail, while the sheep appear vulnerable and agile, their woolly coats catching the light. The composition emphasizes the chase, with the tiger positioned slightly ahead and the sheep trailing behind, creating a sense of motion and urgency. The grassland is lush and green, with scattered wildflowers and a warm, golden glow from the sunlight, enhancing the natural and dramatic atmosphere. Ultra-realistic style, cinematic lighting, and sharp focus bring the scene to life. </think>	

Figure 6: Reasoning trace and generated image of the baseline Bagel model.

Prompts	Thinking Texts	Images
Six cups are arranged in two rows, with three in each row, viewed from above.	*<think> Okay, let's tackle this user prompt. The user wants a prompt for an image of six cups arranged in two rows with three in each row, viewed from above. First, I need to determine the text intention. The user didn't mention any specific text to visualize, so it's "no intention for visualized text."* *Next, I need to structure the prompt with style, content, and aesthetics. The style here is likely "flat design" or "minimalist" since it's a simple arrangement. The content is the six cups in two rows of three, viewed from above. Aesthetics would include color, composition, and atmosphere.* *I should start with the style. "Flat design" is good for simplicity. Then describe the content: six cups arranged in two rows of three, viewed from above. Aesthetics could include color (maybe pastel or neutral), composition (balanced, centered), and atmosphere (clean, modern).* *Check if all elements are included: six cups, two rows of three, viewed from above. No text needed, so no quotes. Ensure the language is concise, no literary fluff. Let me put it all together. </think>*Flat design style, six cups arranged in two rows, three cups in each row, viewed from above. The cups are of different colors: light blue, light green, and light pink, with simple geometric patterns. The composition is centered, the arrangement is neat and orderly, and the color scheme is soft and harmonious, creating a clean and modern atmosphere.	
A boy is holding a piece of wood, and a girl is drawing on it with a pen.	*<think> The model should generate an image where the boy is holding the wood steady, and the girl is actively drawing on it with a pen, emphasizing their collaborative activity. </think>*A highly detailed and realistic scene of a boy and girl working together on a wooden surface, with the boy holding the wood steady with both hands, his posture relaxed yet focused, while the girl sits beside him, actively drawing on the wood with a pen in her right hand. The setting is indoors, with soft, natural lighting illuminating the scene, creating a warm and inviting atmosphere. The wooden surface is smooth and polished, with faint pencil marks visible as the girl draws intricate designs or patterns. The girl appears concentrated, her expression thoughtful, while the boy looks supportive and engaged. The background is minimalistic, featuring a simple wooden table and a few subtle decorative elements, ensuring the focus remains on their collaboration. The image captures a sense of creativity and teamwork, with a slightly artistic and intimate mood.	
A large cat climbing frame has three platforms. On the three platforms, there is a cat, a dog and a mouse lying respectively.	*<think> Okay, let's tackle this user prompt. First, I need to determine if there's a need for visual text. The user didn't mention any specific text to display, so it's "no intention for visualized text."* *Next, I need to rewrite the prompt. The main elements are: a large cat climbing frame with three platforms. On each platform, there's a cat, a dog, and a mouse lying respectively.* *Style: Since it's a scene description, maybe "flat design" or "cartoon illustration" to keep it simple and clear.* *Content: The climbing frame, three platforms, each with a different animal.* *Aesthetic: Colors should be bright and playful, maybe pastel tones. Composition should be centered to show all elements clearly.* *Check if all elements are included: large cat climbing frame, three platforms, cat, dog, mouse on each. No text needed, so no quotes.* *Putting it all together: "Flat design, a large cat climbing frame with three platforms. On the first platform, a cat lies; on the second, a dog lies; on the third, a mouse lies. Pastel color scheme, centered composition, simple and clear illustration style, playful and cute atmosphere."* *</think>*Flat design, a large cat climbing frame with three platforms. On the first platform, a cat lies; on the second, a dog lies; on the third, a mouse lies. Pastel color scheme, centered composition, simple and clear illustration style, playful and cute atmosphere.	
On the grassland, a tiger is chasing two sheep, one big and one small.	*<think> The model should generate an image of a grassland with a tiger in pursuit of two sheep, one larger and one smaller, emphasizing the predator-prey dynamic. </think>*A vast, vibrant grassland under a clear blue sky, featuring a majestic tiger in mid-stride, its powerful muscles taut as it chases two sheep—one large and one small—across the terrain. The scene captures the raw tension of the predator-prey dynamic, with the tiger focused intently on its prey, and the sheep appearing vulnerable yet determined. The grassland is lush and green, with scattered wildflowers and rolling hills in the background, bathed in warm, golden sunlight. The image should feel dynamic and lifelike, with realistic textures and detailed fur patterns on the tiger and sheep, emphasizing the natural beauty and intensity of the moment.	

Figure 7: Reasoning trace and generation result after SFT. During SFT, we standardized the original `<think>` format to match prevalent LLM conventions. Although the model produces extensive reasoning steps, the translation from textual thoughts to fine-grained visual details remains suboptimal.

Prompts	Thinking Texts	Images
Six cups are arranged in two rows, with three in each row, viewed from above.	*<think> The goal is to depict six cups neatly spaced in two horizontal rows of three when seen overhead. Proper alignment and perspective matching are all that change here; no alterations to objects or scene. </think>*A neatly arranged overhead view shows exactly six identical cups evenly spaced in two clear horizontal rows, each containing three cups directlyaligned side by side under uniform lighting for accurate overhead perspective. render each cup as simple ceramic or similar with no clutter nearby to fulfill the specified arrangement from above as described.	
A boy is holding a piece of wood, and a girl is drawing on it with a pen.	*<think> The model appropriately shows the boy gripping wood while the girl writing on it with a pen; accurate; no alterations are necessary. </think>*A young boy gently holds a flat piece of wood horizontally with both hands while a girl skillfully draws designs on the wood's surface using a pen. render scene softly lit indoors with both subjects centered, accurately depicting their actions as described without changing details.	
A large cat climbing frame has three platforms. On the three platforms, there is a cat, a dog and a mouse lying respectively.	*<think> The goal is to depict a cat on one platform, a dog on another, and a mouse on the third with no changes; accurate positioning of each animal on their spots correctly matches the scene described. </think>*A sleek large house cat comfortably lying on its stomach on the middle platform of a spacious pet climbing frame with three distinct circular platforms. Directly to the left on the top platform rests a calm medium-sized fluffy dog, also lying down. On the right of the middle platform, a tiny mouse with smooth fur gently lies on the bottom platform, spaced appropriately. render all three animals peacefully posed facing outward, evenly lit and realistically scaled relative to the climbing structure without altering their positions per the original scene's intent.	
On the grassland, a tiger is chasing two sheep, one big and one small.	*<think> The goal is to depict a tiger chasing the two sheep (large and small) on grassland as described. No alterations; render scene accurately with specified subjects and setting. </think>*A vibrant orange tiger with dark stripes is sprinting to chase a large fluffy white sheep and a smaller white sheep, both calmly running ahead on green grass under clear sky; ensure all accurately portrayed per prompt. render warm realistic textures suitable for scene accuracy without changing intent.	

Figure 8: Reasoning and visual outputs of UniGRPO. Our joint RL optimization produces task-oriented reasoning that guides the synthesis policy toward faithful, photorealistic image generation.

Table 3: Hyperparameters for UniGRPO.

Model Configuration	
Training Timesteps	25
CFG Scale	1
Timestep Shift	3
Image Resolution	1024
Evaluation Timesteps	50
Training	
Group Size	24
Batch Size	32
Reasoning Expert Learning Rate	1e-6
Denoising Expert Learning Rate	3e-5
PPO Epochs	2
Flow Objective Weight (λ)	1
TextGRPO	
KL Divergence Loss Weight	0
Temperature	1
FlowGRPO	
MSE Loss Weight	1.5e-5
Loss Clip Range	1e-6
SDE Window	[0, 5]
SDE Window Size	3
Noise Level	0.8