AdaGRPO: A Capability-Aware Adaptive Enhancement for Flow-based GRPO

Anonymous authors
Paper under double-blind review



Figure 1: **Gallery of AdaGRPO.** By integrating the proposed AdaGRPO, flow models (Flux.1-dev in this figure) experience a substantial leap in the generation performance, yielding remarkable improvements in intricate textures and visual fidelity. **All prompts are listed in the appendix.**

Abstract

Group Relative Policy Optimization (GRPO) has demonstrated remarkable success in aligning text-to-image (T2I) flow models with human preferences. However, we have identified that the learning loop of current flow-based GRPO is fundamentally decoupled from model's current capability, suffering from critical blind spots at both *prompt selection* and *advantage estimation*: (i) existing methods sample prompts randomly, overlooking the substantial impact of data selection on reinforcement learning (RL) efficacy—a factor proven crucial in GRPO for large language models; and (ii) they evaluate sample quality solely relying on intra-group statistics, lacking a global perspective to reliably assess policy progress. To address these issues, we propose **Ada**ptive **GRPO** (**AdaGRPO**), a novel capability-aware RL algorithm tailored for flow models. Specifically, AdaGRPO consists of two principal components: (i) *Online Curriculum Filtering Strategy* dynamically tracks the model's proficiency and adaptively selects prompts that best match its current learning boundary; (ii) *Cross-Level Advantage Fusion* syner-

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.

gistically integrates fine-grained intra-group advantages with global advantages, providing a comprehensive and globally calibrated sample-level evaluation. As a lightweight, plug-and-play module, AdaGRPO can be seamlessly integrated with existing frameworks such as Flow-GRPO, DanceGRPO, and Flow-CPS. Extensive experiments demonstrate that AdaGRPO consistently drives performance gains while significantly stabilizing GRPO training for flow models.

1 INTRODUCTION

Recently, diffusion and flow-based models (Dhariwal & Nichol, 2021; Ho et al., 2020; Podell et al., 2023; Song et al., 2020a;b) have firmly established themselves as the cornerstone of visual generation, exhibiting remarkable proficiency in synthesizing high-quality visual contents (Bu et al., 2025; Labs, 2024; Rombach et al., 2022; Team, 2025; Wan et al., 2025; Zhou et al., 2025a). Despite their impressive generation quality obtained through pre-training on large-scale datasets (Schuhmann et al., 2022; Nan et al., 2024; Chen et al., 2024b), these foundational models often suffer from misalignment with human preferences, such as poor prompt adherence or aesthetic degradation. Consequently, Reinforcement Learning from Human Feedback (RLHF) (Black et al., 2023; Fan et al., 2023; Wang et al., 2026) has become a popular approach for aligning T2I models. By leveraging reward models (Kirstain et al., 2023; Ma et al., 2025; Wang et al., 2026; 2025c; Xu et al., 2023) explicitly designed to encapsulate human intent, RL-based frameworks systematically steer the generation process toward favored visual characteristics and task-specific constraints.

Among various RL techniques (Peng et al., 2025; Rafailov et al., 2023; Schulman et al., 2017; Wallace et al., 2024), Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has recently emerged as a highly promising alternative. By evaluating multiple generated samples for a given prompt and using intra-group comparison to estimate relative advantages, GRPO bypasses the requirement of training a separate value network, making it well-suited for aligning large-scale models. To harness this potential for visual generation, an emerging line of research (Liu et al., 2025; Xue et al., 2025) has adapted GRPO to flow models by replacing deterministic solvers with Stochastic Differential Equations (SDEs), thereby injecting the requisite exploration noise into the sampling trajectory.

Despite these successes, we posit that current flow-based GRPO frameworks (Liu et al., 2025; Xue et al., 2025; He et al., 2025; Zhou et al., 2025b; Li et al., 2025b;a) are fundamentally decoupled from the model's evolving capability during training, suffering from blind spots at two pillars of RL: *prompt selection* ("what data to learn from") and *advantage estimation* ("how to assign credit").

Specifically, regarding *prompt selection*, existing methods usually sample prompts at random. Inspired by the success of prompt selection strategies in reinforcement learning alignment of large language models (LLMs) (Zhang et al., 2025; Yu et al., 2025), we investigate the impact of prompt difficulty on flow-based GRPO. Prior to each training step, we profile all prompts in a candidate batch via their deterministic ODE rewards, then apply filtering heuristics to select which ones actually enter training. As illustrated in Fig. 2 (a), training upon the "easiest" prompts (those yielding the highest rewards) causes severe performance degradation, while employing the "hardest" prompts (the lowest rewards) barely outperforms the random baseline. In contrast, prompts of medium difficulty drive notable gains, corroborating the established finding in LLM alignment that samples of moderate difficulty provide the most useful learning signal (Bae et al., 2026; Cui et al., 2025). However, the median reward of an isolated candidate batch is intrinsically biased, as it is susceptible to divergence from the model's aggregate proficiency. For instance, when an entire batch consists of challenging prompts, the median prompt may still lie beyond the model's capability. The absence of this global perspective also plagues *advantage estimation*. Current methods typically evaluate samples solely via intra-group rewards and thus exhibit severe "myopia". In particular, they erroneously assign positive advantages to subpar samples simply because they are above the local intra-group mean, even if they fall below the model's global capability (false positives), while penalizing high-quality samples that fall below the local mean but actually surpass the global capability (false negatives), as shown in Fig. 2 (b). Without a reliable reference to gauge policy progress, these local biases inevitably obscure the true optimization direction and destabilize training.

To this end, we propose **Ada**ptive **GRPO** (**AdaGRPO**), a novel capability-aware RL algorithm tailored for flow models, addressing the aforementioned blind spots through two principal components. First, *Online Curriculum Filtering Strategy* is introduced to apply prompt selection. Rooted in

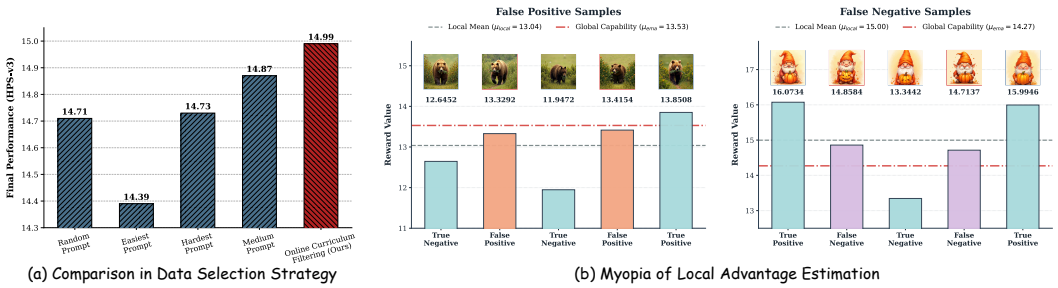


Figure 2: **Observations.** (a) Random sampling (current GRPO methods) and extreme prompts ("Easiest"/"Hardest") yield suboptimal alignment efficacy. While selecting locally moderate prompts ("Medium") offers improvements, it remains biased by the current batch. In contrast, our *Online Curriculum Filtering Strategy* maximizes performance by dynamically identifying moderate tasks through the model's global capability. (b) Relying solely on local intra-group means erroneously produces false positive and false negative advantages that deviate from the model's global capability.

curriculum learning (Soviany et al., 2022), this module maintains an Exponential Moving Average (EMA) of historical rewards to explicitly track the model's global generation proficiency, adaptively selecting candidate prompts at the current learning boundary. This mitigates localized batch bias and ensures a highly constructive optimization landscape. Second, *Cross-Level Advantage Fusion* is proposed to calibrate advantage estimation. By synergistically fusing intra-group local advantages with global advantages, samples are rewarded not only for outperforming their immediate peers but also for surpassing the model's past capability bounds, yielding a faithful signal of policy progress. As a lightweight, plug-and-play module, AdaGRPO seamlessly integrates into prevailing flow-based GRPO frameworks like Flow-GRPO (Liu et al., 2025), DanceGRPO (Xue et al., 2025) and Flow-CPS (Wang & Yu, 2025). Extensive experiments demonstrate that our method consistently drives multi-metric performance gains while significantly stabilizing GRPO training.

Our contributions can be summarized as follows: (i) We identify the structural decoupling in GRPO for flow models, revealing that blind prompt sampling and myopic advantage estimation are bottlenecks causing training instability and suboptimal alignment. (ii) We propose **AdaGRPO**, a novel capability-aware RL algorithm featuring Online Curriculum Filtering Strategy for dynamic data curation and Cross-Level Advantage Fusion for debiased advantage estimation; (iii) AdaGRPO can be seamlessly integrated into diverse existing flow-based GRPO frameworks, offering superior preference alignment and a more stable training process.

2 Related Work

Diffusion and Flow Models. Diffusion models (Ho et al., 2020; Song et al., 2020a;b; Dhariwal & Nichol, 2021) learn to reverse a gradual noising process, enabling high-fidelity visual synthesis across images (Rombach et al., 2022; Podell et al., 2023; Labs, 2024), videos (Guo et al., 2023; Chen et al., 2024a; Wan et al., 2025), and other modalities (Voleti et al., 2024). Flow matching models (Esser et al., 2024; Lipman et al., 2022; Liu et al., 2022) directly learn a continuous-time velocity field along straight-line trajectories between noise and data distributions, offering improved stability and scalability. Leading models such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2023), Flux (Labs, 2024; 2025), Qwen-Image (Wu et al., 2025), CogVideoX (Yang et al., 2024), HunyuanVideo (Kong et al., 2024; Team, 2025), WAN (Wan et al., 2025), and LongCat-Video (Team et al., 2025) have demonstrated remarkable capabilities in generating high-quality visual content.

Alignment for Diffusion and Flow Models. Aligning diffusion/flow models with human preferences has evolved from early PPO-based policy gradients (Black et al., 2023; Schulman et al., 2017; Xu et al., 2023) and DPO variants (Peng et al., 2025; Rafailov et al., 2023; Wallace et al., 2024) toward more efficient online RL frameworks. In particular, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) leverages intra-group relative advantages without a value network, inspiring adaptations to visual generation. Flow-GRPO (Liu et al., 2025) and DanceGRPO (Xue et al., 2025) reformulate deterministic ODE sampling into equivalent SDE trajectories to enable stochastic exploration, establishing the foundational paradigm for flow-based GRPO. Building on it, Flow-CPS (Wang & Yu, 2025) eliminates SDE-induced noise artifacts by strictly aligning noise

injection with the flow-matching scheduler. Subsequent efforts further refine this paradigm from complementary perspectives, such as enhancing training efficiency (Li et al., 2025a; Zheng et al., 2025b), refining reward attribution (Li et al., 2025b; Fu et al., 2025; He et al., 2025; Zhou et al., 2025b; Ling et al., 2026), and enriching reward formulations (Wang et al., 2025b; Bu et al., 2026). Despite these advances, existing methods remain largely oblivious to the dynamic capability of the model. By relying on blind prompt sampling and myopic intra-group advantages, this structural decoupling leads to high training instability and suboptimal alignment efficiency.

Data Selection in Reinforcement Learning. Curriculum learning has long been recognized as an effective strategy to stabilize RL by exposing agents to tasks of progressively increasing difficulty (Bengio et al., 2009; Narvekar et al., 2020). Recent works automate this process by aligning task selection with the agent's evolving capability: ProCuRL (Tzannetos et al., 2023) formalizes the Zone of Proximal Development to maximize learning progress, while Self-Paced RL (Klink et al., 2020) casts sampling as KL-regularized variational inference. In the era of LLM alignment, dynamic sampling techniques (Bae et al., 2026; Cui et al., 2025; Zhang et al., 2025) further refine data curation. Recent efforts have leveraged scoring metrics or group reward dynamics to filter uninformative or zero-variance prompts (Chen et al., 2025; Yu et al., 2025; Zheng et al., 2025a), or formulated online task selection via Bayesian inference and Markov modeling to track solving dynamics (Shen et al., 2025; Mao et al., 2026). AdaGRPO inherits the philosophy of curriculum learning. By dynamically tracking the capability of the model and strategically selecting prompts that reside closest to its current learning boundary, our method effectively smooths the optimization landscape, ensuring robust and stable training progress.

3 AdaGRPO

3.1 Preliminaries

Flow Matching as a Sequential Decision Process. Flow matching (Liu et al., 2022; Lipman et al., 2022) transports samples from a Gaussian prior to a data distribution via a learned velocity field, which can be formulated as a finite-horizon Markov Decision Process (MDP). Given a condition $\mathbf{c}$, the generation trajectory of a flow model is defined as $\Gamma = (\mathbf{s}_T, \mathbf{a}_T, \mathbf{s}_{T-1}, \mathbf{a}_{T-1}, \dots, \mathbf{s}_0, \mathbf{a}_0)$, where each state is $\mathbf{s}_t = (\boldsymbol{x}_t, t, \mathbf{c})$ starting from $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the action $\mathbf{a}_t$ corresponds to the single step denoising process with the policy π_θ. The state transitions follow a deterministic ordinary differential equation (ODE):

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c}). \tag{1}$$

where $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c})$ is the predicted velocity. While this deterministic mapping ensures high-fidelity generation, it inherently lacks the stochasticity required for RL exploration.

ODE-to-SDE Conversion. To adapt flow models for online RL, prior works (Liu et al., 2025; Xue et al., 2025) transform the deterministic ODE into an equivalent Stochastic Differential Equation (SDE). By introducing a diffusion term and compensating the drift, the dynamics become:

$$d\boldsymbol{x}_t = \left(\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c}) + \frac{\sigma_t^2}{2t}\big(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c})\big) \right) dt + \sigma_t d\mathbf{w}_t, \tag{2}$$

where $\mathbf{w}_t$ is the standard Wiener process and $\sigma_t = \eta\sqrt{t/(1-t)}$ governs the magnitude of injected noise with a hyperparameter η. Discretizing this via the Euler–Maruyama scheme over Δt yields:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \left[\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c}) + \frac{\sigma_t^2}{2t}\big(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \mathbf{c})\big) \right] \Delta t + \sigma_t \sqrt{\Delta t}\, \boldsymbol{\epsilon}, \tag{3}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This stochastic formulation injects necessary variance for policy gradient estimation without altering the underlying generative distribution.

GRPO Objective. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a value-free RL paradigm that aligns policies using intra-group feedback. Given a prompt $\mathbf{c}$, the current policy $\pi_{\theta_{\text{old}}}$ samples G trajectories via the SDE, yielding a group of terminal samples $\{\boldsymbol{x}_0^i\}_{i=1}^G$. The advantage for the i-th sample at any timestep t is computed via group-wise reward normalization:

$$\hat{A}_t^i = \frac{R(\boldsymbol{x}_0^i, \mathbf{c}) - \text{mean}\big(\{R(\boldsymbol{x}_0^j, \mathbf{c})\}_{j=1}^G\big)}{\text{std}\big(\{R(\boldsymbol{x}_0^j, \mathbf{c})\}_{j=1}^G\big)}. \tag{4}$$

The policy is then updated by maximizing a clipped surrogate objective with a KL penalty:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\mathbf{c},\{\boldsymbol{x}^i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\left(\min\left(r_t^i(\theta)\hat{A}_t^i, \text{clip}(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t^i\right)\right.\right. \\ \left.\left. - \beta D_{\text{KL}}\left(\pi_\theta(\cdot|\boldsymbol{x}_t, \mathbf{c}) \,\|\, \pi_{\text{ref}}(\cdot|\boldsymbol{x}_t, \mathbf{c})\right)\right)\right], \tag{5}$$

where $r_t^i(\theta) = \pi_\theta(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t^i, \mathbf{c})/\pi_{\theta_{\text{old}}}(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t^i, \mathbf{c})$ is the importance sampling ratio, ε is the clip threshold, β weights the KL penalty against the reference policy π_{ref}.

3.2 Online Curriculum Filtering Strategy

Existing flow-based GRPO methods (Liu et al., 2025; Xue et al., 2025) typically sample training prompts uniformly at random. This blind strategy frequently exposes the policy to extreme tasks that yield either noisy or uninformative optimization signals (Cui et al., 2025; Bae et al., 2026). Furthermore, while applying a localized filtering heuristic (e.g., selecting the median prompt within a candidate batch, as shown in Fig. 2 (a)) can alleviate this issue, it remains biased by the current batch distribution and is disconnected from the model's dynamically evolving capability.

To overcome this, we propose the *Online Curriculum Filtering Strategy*, a lightweight yet effective mechanism rooted in the philosophy of curriculum learning (Soviany et al., 2022), as shown in Fig. 3 (a). Instead of relying on restricted local heuristics, the core idea is to continuously track the model's global generation proficiency and adaptively select candidate prompts that reside at its current learning boundary. Such capability-matched prompts consistently induce the constructive reward variance necessary for meaningful intra-group ranking, providing a clear optimization direction.

Specifically, at each training iteration k, instead of directly performing the SDE group rollout on a random prompt, we first sample a small batch of candidate prompts $\mathcal{B} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_B\}$. For each prompt $\mathbf{c}_b$, we perform a single deterministic ODE sampling to generate a baseline sample $\boldsymbol{x}_0^{b,\text{ODE}} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{c}_b)$, and compute its corresponding reward $R_b^{\text{ODE}} = R(\boldsymbol{x}_0^{b,\text{ODE}}, \mathbf{c}_b)$. To establish a stable capability anchor, we maintain a global historical reward baseline using an Exponential Moving Average (EMA). Let $\mu_{\text{ema}}^{(k)}$ denote the historical mean reward up to iteration k, we update this capability anchor using the mean ODE reward of the current candidate batch:

$$\mu_{\text{ema}}^{(k)} = \alpha\mu_{\text{ema}}^{(k-1)} + (1-\alpha)\frac{1}{B}\sum_{b=1}^{B} R_b^{\text{ODE}}, \tag{6}$$

where $\alpha \in (0,1)$ is the momentum coefficient. Then, we track the EMA variance $(\sigma_{\text{ema}}^{(k)})^2$ to capture the global reward distribution, which is used in *Cross-Level Advantage Fusion* (Section 3.3):

$$(\sigma_{\text{ema}}^{(k)})^2 = \alpha(\sigma_{\text{ema}}^{(k-1)})^2 + (1-\alpha)\frac{1}{B}\sum_{b=1}^{B}\left(R_b^{\text{ODE}} - \mu_{\text{ema}}^{(k)}\right)^2. \tag{7}$$

The EMA mean $\mu_{\text{ema}}^{(k)}$ serves as a robust proxy for the model's current generation capability. For standard single-reward optimization, the candidate batch is filtered to select the prompt $\mathbf{c}_{b^*}$ whose ODE reward is closest to the current capability anchor:

$$b^* = \arg\min_{b\in\{1,\ldots,B\}} \left|R_b^{\text{ODE}} - \mu_{\text{ema}}^{(k)}\right|. \tag{8}$$

Furthermore, this strategy can be seamlessly extended to joint multi-reward optimization involving M distinct reward models, where the optimal prompt is identified by minimizing the sum of normalized deviations across all reward signals:

$$b^* = \arg\min_{b\in\{1,\ldots,B\}} \sum_{m=1}^{M} \frac{\left|R_{b,m}^{\text{ODE}} - \mu_{\text{ema},m}^{(k)}\right|}{\mu_{\text{ema},m}^{(k)}}, \tag{9}$$

where $R_{b,m}^{\text{ODE}}$ and $\mu_{\text{ema},m}^{(k)}$ denote the m-th reward value and its corresponding capability anchor, respectively. The normalization operation eliminates the scale discrepancies among different reward models. Once the optimal prompt $\mathbf{c}_{b^*}$ is selected, we execute the full stochastic group rollout (size G) exclusively on $\mathbf{c}_{b^*}$ for subsequent optimization.

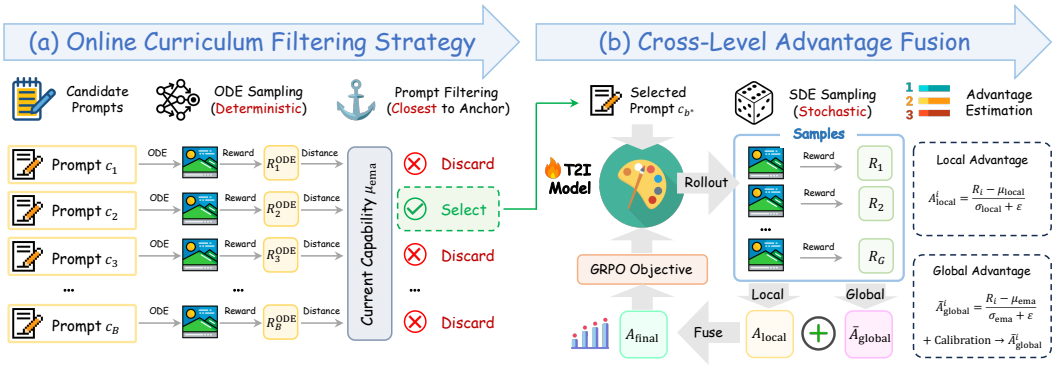


Figure 3: **Pipeline of AdaGRPO**. (a) First, *Online Curriculum Filtering Strategy* evaluates candidate prompts via deterministic ODE sampling and adaptively selects the one that best matches the model's current capability anchor (μ_{ema}). The selected prompt is utilized for stochastic SDE rollout. (b) Then, *Cross-Level Advantage Fusion* integrates the intra-group local advantage with the history-calibrated global advantage to formulate a comprehensive signal (A_{final}) for GRPO optimization.

3.3 Cross-Level Advantage Fusion

While the standard GRPO effectively avoids the need for a separate value network, its exclusive reliance on intra-group evaluation restricts its field of view. As demonstrated in Fig. 2 (b), this localized paradigm generates false positive advantages by rewarding subpar samples simply because they exceed the intra-group mean, while producing false negative advantages by penalizing high-quality samples that happen to fall below their immediate peers but actually surpass the global capability. Such miscalibrated sample-level evaluations hinder an accurate assessment of policy progress. To this end, we propose *Cross-Level Advantage Fusion*, which synergistically integrates fine-grained local rankings with a global capability baseline, as depicted in Fig. 3 (b).

Local Advantage Estimation. Given the selected prompt $\mathbf{c}_{b^*}$ and its generated group of G samples $\{\boldsymbol{x}_0^i\}_{i=1}^G$, we first compute the standard intra-group local advantage (timestep t is omitted for brevity):

$$A^i_{\text{local}} = \frac{R_i - \mu_{\text{local}}}{\sigma_{\text{local}} + \epsilon}, \tag{10}$$

where R_i is the reward of the i-th SDE sample generated with $\mathbf{c}_{b^*}$, μ_{local} and σ_{local} are the mean and standard deviation of $\{R_i\}_{i=1}^G$, respectively, and ϵ is a small constant for numerical stability.

Global Advantage Calibration. To inject a global perspective, we leverage the historical reward statistics, $\mu^{(k)}_{\text{ema}}$ and $\sigma^{(k)}_{\text{ema}}$, maintained in Section 3.2, to derive a raw global advantage:

$$\tilde{A}^i_{\text{global}} = \frac{R_i - \mu^{(k)}_{\text{ema}}}{\sigma^{(k)}_{\text{ema}} + \epsilon}. \tag{11}$$

However, directly utilizing $\tilde{A}^i_{\text{global}}$ breaks the critical zero-mean property of GRPO, potentially destabilizing the policy optimization. To enforce a strict zero-mean distribution while preserving the sign of each sample's absolute progression, we introduce a conditional sign-preserving normalization step. Let $\mathcal{P} = \{i \mid \tilde{A}^i_{\text{global}} > 0\}$ and $\mathcal{N} = \{i \mid \tilde{A}^i_{\text{global}} < 0\}$ denote the indices of positive and negative global advantages, respectively. They are scaled conditionally as follows:

$$\bar{A}^i_{\text{global}} = \begin{cases} \dfrac{\tilde{A}^i_{\text{global}}}{\sum_{j\in\mathcal{P}} \tilde{A}^j_{\text{global}}}, & \text{if } i \in \mathcal{P} \text{ and } \mathcal{P}, \mathcal{N} \neq \emptyset \\ \dfrac{\tilde{A}^i_{\text{global}}}{\sum_{j\in\mathcal{N}} |\tilde{A}^j_{\text{global}}|}, & \text{if } i \in \mathcal{N} \text{ and } \mathcal{P}, \mathcal{N} \neq \emptyset \\ \tilde{A}^i_{\text{global}} - \frac{1}{G}\sum_{j=1}^{G} \tilde{A}^j_{\text{global}}, & \text{otherwise.} \end{cases} \tag{12}$$

In the standard case (both sets non-empty), this operation scales the sum of positive terms to 1 and negative terms to -1, guaranteeing $\sum_i \bar{A}^i_{\text{global}} = 0$. In the rare event of a unilateral batch (i.e., $\mathcal{P} = \emptyset$ or $\mathcal{N} = \emptyset$), we dynamically fall back to standard mean-centering to prioritize training stability.

Cross-Level Advantage Fusion. Finally, we formulate the comprehensive advantage signal by aggregating the local and global advantages:

$$A^i_{\text{final}} = A^i_{\text{local}} + \bar{A}^i_{\text{global}}. \tag{13}$$

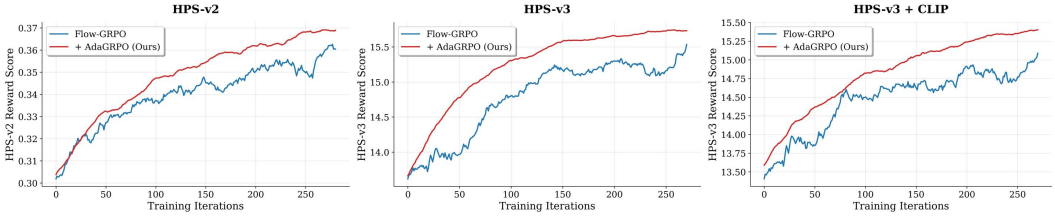


Figure 4: **Reward Curves during Training.** The proposed AdaGRPO facilitates significantly smoother training dynamics and higher performance ceilings across diverse training configurations.

By replacing the standard advantage $\hat{A}_t^i$ in Equation 5 with this fused advantage A_{final}^i, AdaGRPO ensures that samples are rewarded not only for outperforming their peers but also for surpassing the model's historical capability bounds, providing a calibrated gradient direction. The overview of our AdaGRPO framework is illustrated in Fig. 3. The algorithm is provided in Section B in the appendix.

4 Experiments

4.1 Implementation Details

Datasets and Models. Following prior works (Xue et al., 2025; Li et al., 2025a; Zhou et al., 2025b), the HPD (Wu et al., 2023) dataset is utilized as the prompt dataset. This comprehensive corpus supplies over 100K diverse prompts to drive the RL training, alongside a distinct set of 400 prompts for evaluation. For the generative backbone, all experiments in this section are built upon Flux.1-dev (Labs, 2024), one of the most capable open-sourced flow models. Further details and results on another T2I backbone are provided in Section C and Section F in the appendix, respectively.

Baselines. To demonstrate its architecture-agnostic nature, we implement AdaGRPO upon three representative flow-based GRPO baselines: Flow-GRPO (Liu et al., 2025), DanceGRPO (Xue et al., 2025), and Flow-CPS (Wang & Yu, 2025). To improve training efficiency, the few-step training mechanism of Flow-GRPO-Fast (Liu et al., 2025) is adopted by all assessed methods.

Evaluation Metrics. For a comprehensive assessment, we assemble a diverse suite of automated evaluation metrics that capture different facets of generation quality, including (i) *CLIP/BLIP-based Reward Models*: HPS-v2 (Wu et al., 2023), CLIP (Radford et al., 2021) and ImageReward (**IR**) (Xu et al., 2023); (ii) *LVLM-based Reward Models*: HPS-v3 (Ma et al., 2025) and UnifiedReward-v1/v2 (**UR-v1/v2**) (Wang et al., 2025c); and (iii) *General T2I Benchmarks*: UniGenBench (Wang et al., 2025a), a unified and versatile benchmark for image generation. This comprehensive benchmark covers ten distinct categories that span essential aspects such as conceptual fidelity, visual appeal, and text-image alignment, offering a holistic measure of generative capability.

Training Paradigms. Following previous studies (Xue et al., 2025; Li et al., 2025a; Zhou et al., 2025b), we train AdaGRPO under two distinct configurations: (i) *Single Reward*: the model is fine-tuned using a solitary reward signal (either HPS-v2 or HPS-v3); (ii) *Multi-Reward*: the policy is jointly optimized under signals from both HPS-v3 and CLIP for more robust and generalizable alignment outcomes. Training results on UnifiedReward-v2 are shown in Section F in the appendix.

Sampling Details. Following previous works (Xue et al., 2025; Zhou et al., 2025b), a group size of $G = 12$ is adopted and the total number of sampling steps is configured as $T = 16$. The candidate prompt batch size B and the momentum coefficient α are set to 10 and 0.6, respectively.

Optimization Details. All experiments are conducted on $8\times$ NVIDIA H200 GPUs with the batch size set to 1. The AdamW optimizer is utilized with a learning rate of 2×10^{-6} and a weight decay of 1×10^{-4}. For efficiency, `bfloat16` (bf16) mixed-precision is leveraged during training.

4.2 Main Results

Quantitative Evaluation. The quantitative assessments are presented in Tab. 1 and Tab. 2. Under both single reward (HPS-v2/v3) and multi-reward (HPS-v3 + CLIP) settings, AdaGRPO consistently brings substantial improvements to the prevailing baselines (Flow-GRPO, DanceGRPO, and Flow-CPS), validating its effectiveness and architecture-agnostic nature. Specifically, as shown in Tab. 1,

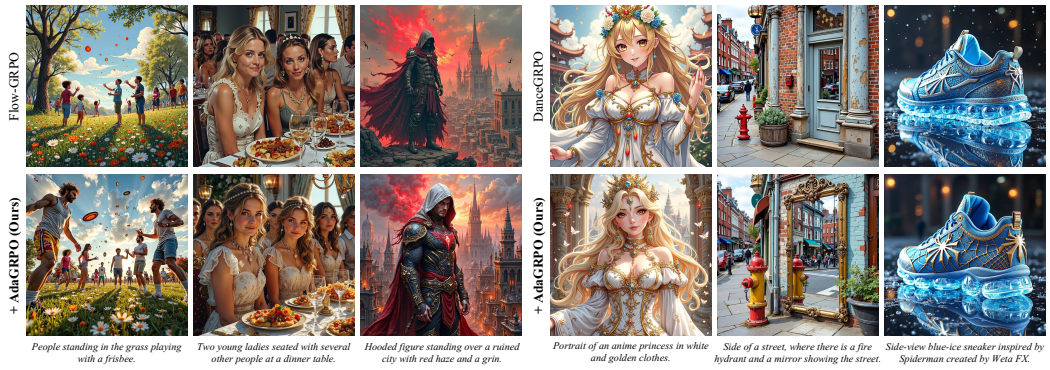


Figure 5: **Qualitative Comparisons with Baselines on HPS-v2.** Best viewed zoomed in.

Table 1: Quantitative comparison across different settings and frameworks. **Bold** values indicate the best result within each pair. Shaded rows denote results with our AdaGRPO. UR-v2-A, UR-v2-C and UR-v2-S represent the Alignment, Coherence and Style dimensions of UR-v2, respectively.

Reward Model	Method	HPS-v2↑	HPS-v3↑	UR-v2-A↑	UR-v2-C↑	UR-v2-S↑	CLIP↑	IR↑	UR-v1↑
/	Flux.1-dev	0.3065	13.2803	3.2550	3.6547	3.2333	0.3901	1.0546	3.6006
HPS-v2	Flow-GRPO	0.3463	14.4374	**3.2293**	3.6236	3.3395	**0.3754**	1.2435	3.5167
	+ AdaGRPO	**0.3558**	**14.8870**	3.2222	**3.6595**	**3.3840**	0.3676	**1.2645**	**3.5285**
	DanceGRPO	0.3430	14.4687	**3.2291**	3.6253	3.3089	**0.3724**	1.2178	3.5129
	+ AdaGRPO	**0.3516**	**14.5155**	3.2209	**3.6584**	**3.3891**	0.3685	**1.2539**	**3.5355**
	Flow-CPS	0.3492	14.8044	**3.2387**	3.6660	3.3679	**0.3745**	1.2561	3.5276
	+ AdaGRPO	**0.3579**	**14.9521**	3.2264	**3.6676**	**3.3722**	0.3691	**1.2575**	**3.5537**
HPS-v3	Flow-GRPO	0.3272	14.7143	**3.2300**	3.6728	3.2836	**0.3787**	1.1475	**3.5821**
	+ AdaGRPO	**0.3319**	**15.0982**	3.2018	**3.6774**	**3.3154**	0.3699	**1.1490**	3.5512
	DanceGRPO	0.3275	14.6675	**3.2345**	3.6450	3.2843	**0.3813**	1.1431	**3.5460**
	+ AdaGRPO	**0.3301**	**14.9501**	3.2266	**3.6724**	**3.3052**	0.3780	**1.1487**	3.5376
	Flow-CPS	0.3291	14.7578	**3.2312**	3.6550	3.2786	**0.3783**	**1.1598**	**3.5375**
	+ AdaGRPO	**0.3311**	**15.1093**	3.2169	**3.6823**	**3.3133**	0.3734	1.1548	3.5193
HPS-v3 + CLIP	Flow-GRPO	0.3250	14.2459	**3.2605**	3.6231	3.2487	0.3938	1.1988	3.6260
	+ AdaGRPO	**0.3288**	**14.6114**	3.2566	**3.6418**	**3.2511**	**0.3952**	**1.2018**	**3.6340**
	DanceGRPO	0.3189	13.8573	3.2483	3.6294	3.2499	0.3914	1.1406	3.5859
	+ AdaGRPO	**0.3246**	**14.3156**	**3.2514**	**3.6357**	**3.2500**	**0.3924**	**1.2216**	**3.6203**
	Flow-CPS	0.3252	14.3248	3.2608	3.6180	3.2358	0.3926	1.2222	3.6339
	+ AdaGRPO	**0.3286**	**14.6044**	**3.2642**	**3.6338**	**3.2475**	**0.3942**	**1.2300**	**3.6419**

AdaGRPO delivers the best performance on the majority of evaluation metrics, with particularly notable gains in HPS-related scores, coherence (UR-v2-C), style (UR-v2-S) and ImageReward. Meanwhile, when incorporating the CLIP reward model to enforce semantic alignment during joint multi-reward training, AdaGRPO achieves consistent improvements across nearly all evaluation dimensions. Furthermore, as detailed in Tab. 2, the comprehensive evaluation on UniGenBench corroborates our superiority in fine-grained visual synthesis. The training reward curves for Flow-GRPO (with or without our AdaGRPO) are presented in Fig. 4.

Qualitative Comparison. As depicted in Fig. 5 and Fig. 6, AdaGRPO consistently surpasses the baselines in visual fidelity, aesthetic appeal, and semantic adherence. In the *"dinner table"* and *"man finger nose"* cases, our method renders portraits with significantly more natural skin textures, precise facial anatomy, and cinematic lighting gradients, overcoming the plastic appearance generated by baselines. For the *"warrior"* and *"blue-ice sneaker"* examples, AdaGRPO substantially enhances the visual richness by synthesizing intricate ornamental details, vivid glowing elements, and realistic material reflections. Furthermore, our method exhibits robust adherence to complex spatial compositions. In the *"mirror"* case, the baseline completely ignores the specified reflective element, whereas AdaGRPO accurately generates an ornate mirror reflecting the street scene. Similarly, in the *"frisbee"* case, it transforms a static and chaotic baseline generation into a highly dynamic action composition with a more immersive atmosphere. More results are provided in Section D in appendix.

4.3 Ablation and Analysis

We conducted ablation studies of AdaGRPO on Flow-GRPO framework under the HPS-v2 training setting. Given that the two proposed components build upon one another, we begin by ablating the



Figure 6: **Qualitative Comparisons with Baselines on HPS-v3.** Best viewed zoomed in.

Table 2: Quantitative comparison on UniGenBench (Training Setting: HPS-v3 + CLIP). **Bold** values indicate the best result within each pair. Shaded rows denote results with our AdaGRPO.

Method	Overall↑	Style↑	Know.↑	Attri.↑	Act.↑	Rel.↑	Comp.↑	Gram.↑	Reason.↑	Layout↑	Text↑
Flux.1-dev	59.86	84.60	85.76	64.32	61.22	66.37	46.26	58.69	27.06	70.71	33.62
Flow-GRPO	62.65	85.80	89.08	70.19	66.06	69.67	51.03	**60.70**	29.36	73.88	**30.75**
+ AdaGRPO	**63.24**	**86.40**	**89.56**	**70.73**	**66.83**	**71.70**	**51.16**	60.29	**30.28**	**75.00**	30.46
DanceGRPO	61.15	85.00	88.29	68.48	61.12	**70.05**	48.20	59.36	28.90	71.64	**30.46**
+ AdaGRPO	**62.41**	**86.50**	**90.51**	**69.02**	**62.93**	69.42	**51.68**	**60.56**	**30.50**	**73.13**	29.89
Flow-CPS	62.65	86.10	90.03	69.66	**65.87**	71.32	52.19	**61.36**	29.13	73.51	27.30
+ AdaGRPO	**63.31**	**86.50**	**90.66**	**69.98**	65.11	**72.84**	**52.71**	60.70	**29.82**	**74.07**	**30.75**

Online Curriculum Filtering Strategy, and subsequently investigate the impact of incorporating the Cross-Level Advantage Fusion under its optimal configuration. The results are presented in Tab. 3.

Effects of Online Curriculum Filtering Strategy. This strategy relies on two hyperparameters: the momentum coefficient α and the candidate batch size B. α controls the update rate of the historical capability anchor. As depicted in Tab. 3 (a), a moderate $\alpha = 0.6$ achieves the best HPS-related scores with competitive UR-v2 performance (averaged over the Alignment, Coherence and Style dimensions), effectively balancing long-term historical knowledge with current batch statistics. Meanwhile, B defines the search space for prompt selection. While a larger B theoretically enables more precise capability matching, Tab. 3 (b) reveals a clear diminishing return when scaling B beyond 10. Considering the computational overhead of additional ODE sampling (Section G in the appendix), $B = 10$ is chosen to strike a balance between training efficiency and alignment performance.

Effects of Cross-Level Advantage Fusion. As shown in Tab. 3 (c), relying solely on intra-group evaluation traps the policy in local optima, yielding suboptimal performance. By contrast, integrating the historical baseline significantly elevates both HPS-related scores and the averaged UR-v2 metrics, confirming that our cross-level fusion effectively rectifies myopic local biases and drives effective policy progress.

Table 3: Ablation study results. **Bold** and underlined indicate the best and second-best results, respectively.

Components	Choices	HPS-v2↑	HPS-v3↑	UR-v2 (averaged)↑
(a) Momentum Coefficient α	0.9	0.3490	14.7034	3.4107
	0.8	0.3493	14.6675	3.4212
	0.7	0.3499	14.6815	**3.4241**
	0.6	**0.3507**	**14.7761**	3.4197
	0.5	0.3503	14.7340	3.4175
(b) Candidate Batch Size B	5	0.3480	14.6277	3.4135
	10	0.3507	14.7761	3.4197
	20	**0.3510**	**14.8026**	**3.4199**
(c) Cross-Level Adv Fusion	w/o	0.3507	14.7761	3.4197
	w	**0.3558**	**14.8870**	**3.4219**

5 Conclusion

In this paper, we identify that blind prompt selection and myopic advantage estimation in current flow-based GRPO lead to training instability and suboptimal alignment. To address this, we propose AdaGRPO, a lightweight capability-aware RL framework. It features *Online Curriculum Filtering Strategy* to dynamically match training prompts with the model's evolving capability, and *Cross-Level Advantage Fusion* to integrate local rankings with a global baseline for calibrated sample-level evaluation. Extensive experiments demonstrate that AdaGRPO seamlessly integrates into prevailing architectures, consistently driving superior generation quality and highly stable training dynamics.

AI USE STATEMENT

In this paper, we used generative AI tools only for minor language polishing. We take responsibility for the content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available datasets, models, and reward models, and all experiments follow standard community practice. We do not foresee specific societal risks or ethical concerns beyond those commonly associated with text-to-image generation.

REPRODUCIBILITY STATEMENT

We strive to make our experimental results fully verifiable and accessible to the broader academic community. To facilitate reproducibility, we provide comprehensive experimental details and configurations throughout the main paper and appendix, including implementation settings, training hyperparameters, and evaluation protocols. In addition, the source code of AdaGRPO will be publicly released. At this stage, an anonymized implementation is provided in the supplementary material (described in Section H of the appendix). We sincerely hope that these resources will serve as a robust baseline and facilitate future research on reinforcement learning and flow model alignment.

REFERENCES

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 700–719, 2026.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

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

Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025.

Jiazi Bu, Pengyang Ling, Yujie Zhou, Yibin Wang, Yuhang Zang, Tianyi Wei, Xiaohang Zhan, Jiaqi Wang, Tong Wu, Xingang Pan, et al. From sparse to dense: Multi-view grpo for flow models via augmented condition space. *arXiv preprint arXiv:2603.12648*, 2026.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Zhuoyue Chen, Jihai Zhang, Ben Liu, Fangquan Lin, and Wotao Yin. Scale down to speed up: Dynamic data selection for reinforcement learning. *Training*, 2500:3000, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

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. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Xiaolong Fu, Lichen Ma, Zipeng Guo, Gaojing Zhou, Chongxiao Wang, ShiPing Dong, Shizhe Zhou, Ximan Liu, Jingling Fu, Tan Lit Sin, et al. Dynamic-treerpo: Breaking the independent trajectory bottleneck with structured sampling. *arXiv preprint arXiv:2509.23352*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Pascal Klink, Carlo D'Eramo, Jan R Peters, and Joni Pajarinen. Self-paced deep reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9216–9227, 2020.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025b.

Pengyang Ling, Jiazi Bu, Yujie Zhou, Yibin Wang, Zhenyu Hu, Zihan Zhang, Yi Jin, Huaian Chen, and Yuhang Zang. Pave-grpo: Beyond instantaneous guidance through principled average velocity decomposition. *arXiv preprint arXiv:2606.01636*, 2026.

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.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Yixiu Mao, Yun Qu, Qi Wang, Heming Zou, and Xiangyang Ji. Dynamics-predictive sampling for active rl finetuning of large reasoning models. *arXiv preprint arXiv:2603.10887*, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

Liang Peng, Boxi Wu, Haoran Cheng, Yibo Zhao, and Xiaofei He. Sudo: Enhancing text-to-image diffusion models with self-supervised direct preference optimization. *arXiv preprint arXiv:2504.14534*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, 36:53728–53741, 2023.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

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

Qianli Shen, Daoyuan Chen, Yilun Huang, Zhenqing Ling, Yaliang Li, Bolin Ding, and Jingren Zhou. Bots: A unified framework for bayesian online task selection in llm reinforcement finetuning. *arXiv preprint arXiv:2510.26374*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL `https://arxiv.org/abs/2511.18870`.

Georgios Tzannetos, Bárbara Gomes Ribeiro, Parameswaran Kamalaruban, and Adish Singla. Proximal curriculum for reinforcement learning agents. *arXiv preprint arXiv:2304.12877*, 2023.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, 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.

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

Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

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

Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *arXiv preprint arXiv:2602.02380*, 2026.

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*, 2025.

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

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*, 36:15903–15935, 2023.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. *arXiv preprint arXiv:2504.14286*, 2025.

Haizhong Zheng, Yang Zhou, Brian R Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *arXiv preprint arXiv:2506.02177*, 2025a.

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. *arXiv preprint arXiv:2509.16117*, 2025b.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13315–13325, October 2025a.

Yujie Zhou, Pengyang Ling, Jiazi Bu, Yibin Wang, Yuhang Zang, Jiaqi Wang, Li Niu, and Guangtao Zhai. G2rpo: Granular grpo for precise reward in flow models. *arXiv preprint arXiv:2510.01982*, 3, 2025b.

A APPENDIX

In the appendix, we present the following additional materials to complement the main paper:

- The algorithm of AdaGRPO (Section B).
- Additional implementation details (Section C).
- Additional qualitative samples (Section D).
- Text prompts for image generation in both the main paper and appendix (Section E).
- Additional experimental results (Section F).
- The limitation of our method (Section G).
- The description of supplementary material (Section H).

B ALGORITHM

The overall training procedure of the proposed AdaGRPO is detailed in Algorithm 1. At each training iteration, the *Online Curriculum Filtering Strategy* first profiles a candidate prompt batch via deterministic ODE sampling, updates the capability anchor $(\mu_{\text{ema}}, \sigma_{\text{ema}})$, and selects the prompt residing closest to the model's current learning boundary. Then, a stochastic SDE group rollout is performed on the selected prompt, and the *Cross-Level Advantage Fusion* integrates the intra-group local advantage with the history-calibrated global advantage to produce a comprehensive signal A_{final}, which is finally substituted into the GRPO objective (Eq. 5) for policy optimization.

Algorithm 1 Adaptive Group Relative Policy Optimization (AdaGRPO)

Require: Prompt dataset $\mathcal{D}$, policy model π_θ, reward model R, group size G, candidate batch size B, EMA momentum α, total sampling steps T, SDE sampling timestep set M, learning rate λ, total training iterations E
for training iteration $k = 1$ **to** E **do**
  Update old policy model: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  # 1. Online Curriculum Filtering Strategy
  Sample a candidate prompt batch $\mathcal{B} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_B\} \sim \mathcal{D}$
  for candidate prompt $\mathbf{c}_b \in \mathcal{B}$ **do**
    Generate baseline sample $\boldsymbol{x}_0^{b,\text{ODE}} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{c}_b)$ via deterministic ODE sampling
    Compute reward $R_b^{\text{ODE}} = R(\boldsymbol{x}_0^{b,\text{ODE}}, \mathbf{c}_b)$
  end for
  Update capability anchors $\mu_{\text{ema}}^{(k)}$ and $\sigma_{\text{ema}}^{(k)}$ via EMA — *Eq. 6 and Eq. 7*
  Select the optimal prompt $\mathbf{c}_{b^*}$ closest to the capability anchor — *Eq. 8 or Eq. 9 (for multi-reward)*
  # 2. Stochastic SDE Group Rollout on the Selected Prompt
  Initialize noise: $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$
  for $t = T$ **down to** 1 **do**
    if $t \in M$ **then**
      SDE Sampling: generate $\boldsymbol{x}_{t-1}^i$ for $i = 1 \ldots G$
    else
      ODE Sampling: generate $\boldsymbol{x}_{t-1}$
    end if
  end for
  Obtain a group of samples $\{\boldsymbol{x}_0^i\}_{i=1}^G$ and compute rewards $\{R_i\}_{i=1}^G$ with $\mathbf{c}_{b^*}$
  # 3. Cross-Level Advantage Fusion
  Compute intra-group local advantages A_{local}^i — *Eq. 10*
  Compute raw global advantages $\tilde{A}_{\text{global}}^i$ — *Eq. 11*
  Apply conditional sign-preserving normalization $\rightarrow \bar{A}_{\text{global}}^i$ — *Eq. 12*
  Fuse advantages: $A_{\text{final}}^i = A_{\text{local}}^i + \bar{A}_{\text{global}}^i$ — *Eq. 13*
  # 4. Policy Update
  Compute the GRPO objective $\mathcal{J}_{\text{GRPO}}(\theta)$ with A_{final}^i — *Eq. 5*
  Update policy: $\theta \leftarrow \theta + \lambda \nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta)$
end for

C Additional Implementation Details

Tab. 4 presents the detailed hyperparameter settings used in our study, which were kept consistent throughout all experiments unless otherwise specified.

Table 4: Hyperparameter settings in our experiments.

Parameter	Value	Parameter	Value
Random seed	42	Learning rate	2×10^{-6}
Train batch size	1	Weight decay	1×10^{-4}
Warmup steps	0	Mixed precision	`bfloat16`
Dataloader workers	4	Max grad norm	1.0
The number of GPUs	8	Eta	0.7
Group size	12	Scheduler shift	3
Sampling steps	16	Adv. clip max	5.0
Init same noise	Yes	SDE timestep set	$\{0, 2, 4, 6\}$

D Additional Qualitative Samples

In this section, we provide additional qualitative comparisons between the proposed AdaGRPO and baseline methods, as shown in Fig. 8, Fig. 9, Fig. 10, and Fig. 11. We also present more visual results of AdaGRPO in Fig. 12, Fig. 13, Fig. 14, and Fig. 15, as well as generated results obtained using the same prompts but different random seeds in Fig. 16 and Fig. 17.

E Text Prompts

Text prompts used to generate images in this paper are listed in Tab. 8, Tab. 9 and Tab. 10.

F Additional Experimental Results

F.1 Additional Experiments with UnifiedReward-v2

To further validate the versatility of our proposed AdaGRPO, we conduct additional experiments utilizing UnifiedReward-v2 (UR-v2) (Wang et al., 2025c) as the reward model. Unlike CLIP or HPS variants, UR-v2 is a state-of-the-art LVLM-based reward model that provides comprehensive, multi-dimensional evaluations encompassing Alignment (UR-v2-A), Coherence (UR-v2-C), and Style (UR-v2-S). Specifically, we integrate AdaGRPO into the three representative baseline frameworks (Flow-GRPO, DanceGRPO, and Flow-CPS) and train them using the averaged UR-v2 score. As shown in Tab. 5, AdaGRPO consistently outperforms the standard GRPO baselines across all architectures, achieving superior averaged UR-v2 scores while exhibiting robust generalization to unseen auxiliary metrics such as HPS-v2/v3, UR-v1, and ImageReward (IR).

Table 5: Quantitative comparison of models trained with UnifiedReward-v2 (UR-v2). **Bold** values indicate the best result within each pair. Shaded rows denote results with our AdaGRPO.

Reward Model	Method	HPS-v2↑	HPS-v3↑	UR-v2-A↑	UR-v2-C↑	UR-v2-S↑	CLIP↑	IR↑	UR-v1↑
/	Flux.1-dev	0.3065	13.2803	3.2550	3.6547	3.2333	0.3901	1.0546	3.6006
UR-v2	Flow-GRPO	0.3173	13.8476	**3.2390**	3.6973	3.3392	**0.3823**	1.1127	3.5376
	+ AdaGRPO	**0.3202**	**13.9199**	3.2320	**3.7358**	**3.3670**	0.3771	**1.1159**	**3.5804**
	DanceGRPO	0.3202	14.0876	3.2336	3.6807	3.3162	0.3802	1.0962	3.5181
	+ AdaGRPO	**0.3210**	**14.1153**	**3.2458**	**3.7058**	**3.3297**	**0.3836**	**1.1387**	**3.5980**
	Flow-CPS	0.3213	14.1655	**3.2406**	3.7146	3.3794	**0.3786**	1.1246	3.5639
	+ AdaGRPO	**0.3265**	**14.1743**	3.2382	**3.7272**	**3.3925**	0.3774	**1.1452**	**3.5657**



Figure 7: **Qualitative Comparisons on SD3.5-Medium.** Reward Model: HPS-v3.

Table 6: Quantitative comparison on SD3.5-Medium with HPS-v3 as the reward model. **Bold** values indicate the best result within each pair. Shaded rows denote results with our AdaGRPO.

Reward Model	Method	HPS-v2↑	HPS-v3↑	UR-v2-A↑	UR-v2-C↑	UR-v2-S↑	CLIP↑	IR↑	UR-v1↑
/	SD3.5-Medium	0.2901	10.8301	3.2213	3.5073	3.1067	0.4051	1.0049	3.5909
HPS-v3	FlowGRPO	0.3339	13.4720	3.1995	3.5395	3.2192	**0.3877**	1.2501	3.4633
	+ AdaGRPO	**0.3383**	**13.9062**	**3.2236**	**3.5859**	**3.2407**	0.3854	**1.2919**	**3.5671**

F.2 Additional Experiments on StableDiffusion 3.5

To verify the generalization capability of AdaGRPO across different backbones, we conduct additional experiments on SD3.5-Medium (Esser et al., 2024), another open-sourced T2I flow model. Specifically, we integrate AdaGRPO into the SD3.5 + Flow-GRPO (Liu et al., 2025) training setting. As shown in Tab. 6, AdaGRPO achieves remarkable improvements over the Flow-GRPO baseline across the majority of evaluation metrics, demonstrating that our capability-aware design generalizes seamlessly to other backbones. Visual comparisons are presented in Fig. 7.

G Limitation and Discussion

While AdaGRPO demonstrates superior performance and training stability, it faces certain constraints. Similar to dynamic data sampling strategies for LLM (Yu et al., 2025; Bae et al., 2026), our online prompt filtering inevitably introduces some overhead. This overhead mainly comes from deterministic ODE profiling over candidate prompts, which can be efficiently parallelized given the moderate VRAM requirements of T2I models. As shown in Tab. 7, the default setting ($B = 10$) introduces only a 21.0% increase in runtime while maintaining nearly constant memory usage. Future work could explore more efficient filtering strategies, such as employing low-bit models (e.g., `int4`) during profiling while reserving full-precision models (e.g., `fp16`) for SDE rollouts.

Table 7: Computational overhead of prompt filtering with different candidate batch sizes.

B	Iter. Time (s)	Memory (GB)
1	34.617 (+0.0%)	63.85
5	38.197 (+10.3%)	63.85
10	**41.877** (+21.0%)	**63.87**
20	51.304 (+48.2%)	63.88

H Source Code

We provide the code implementation (for Flux) of AdaGRPO for reference in the zip archive included in the supplementary material.





Figure 8: **Additional Comparison Results on HPS-v2. (1/2)**

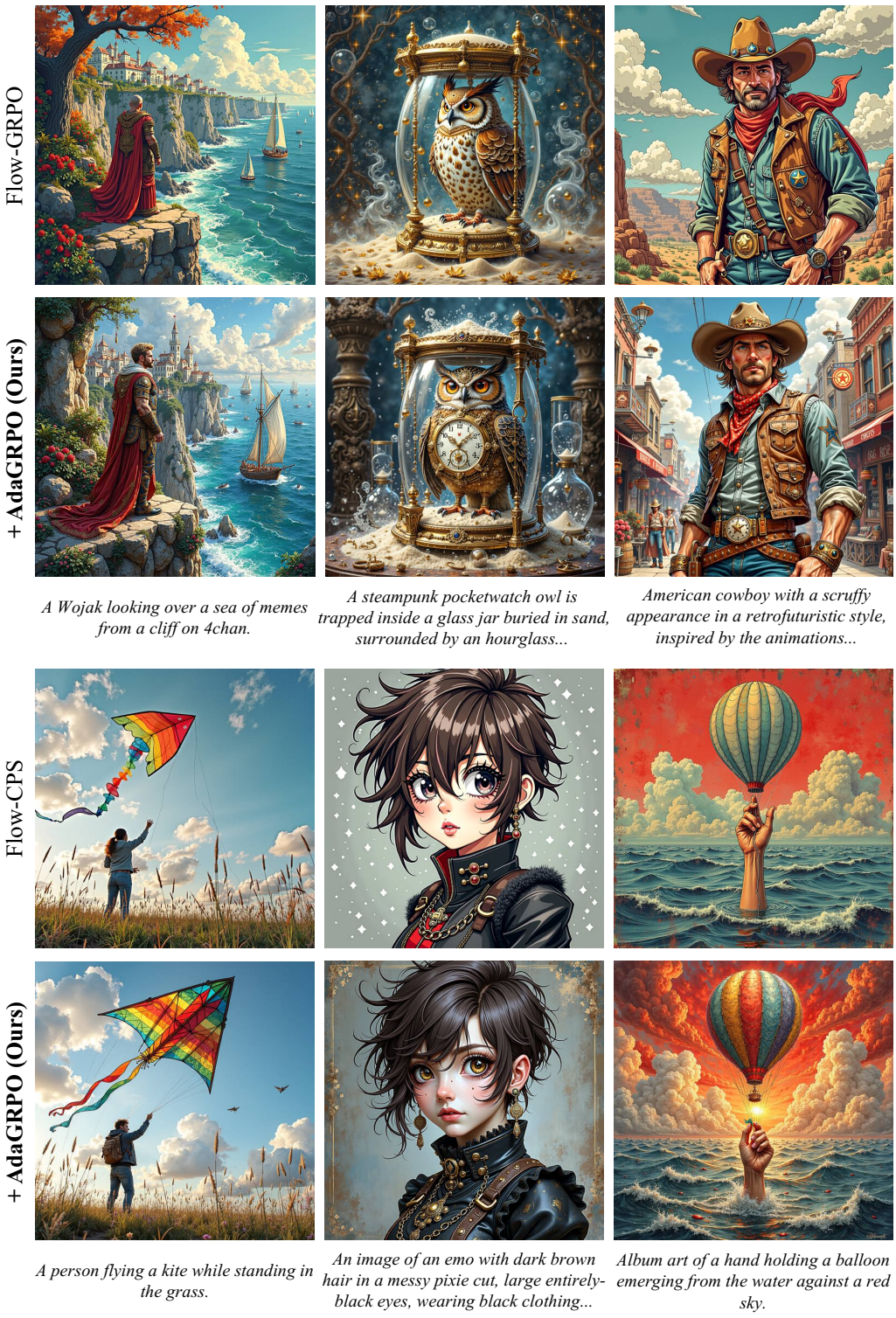


Figure 9: **Additional Comparison Results on HPS-v2. (2/2)**

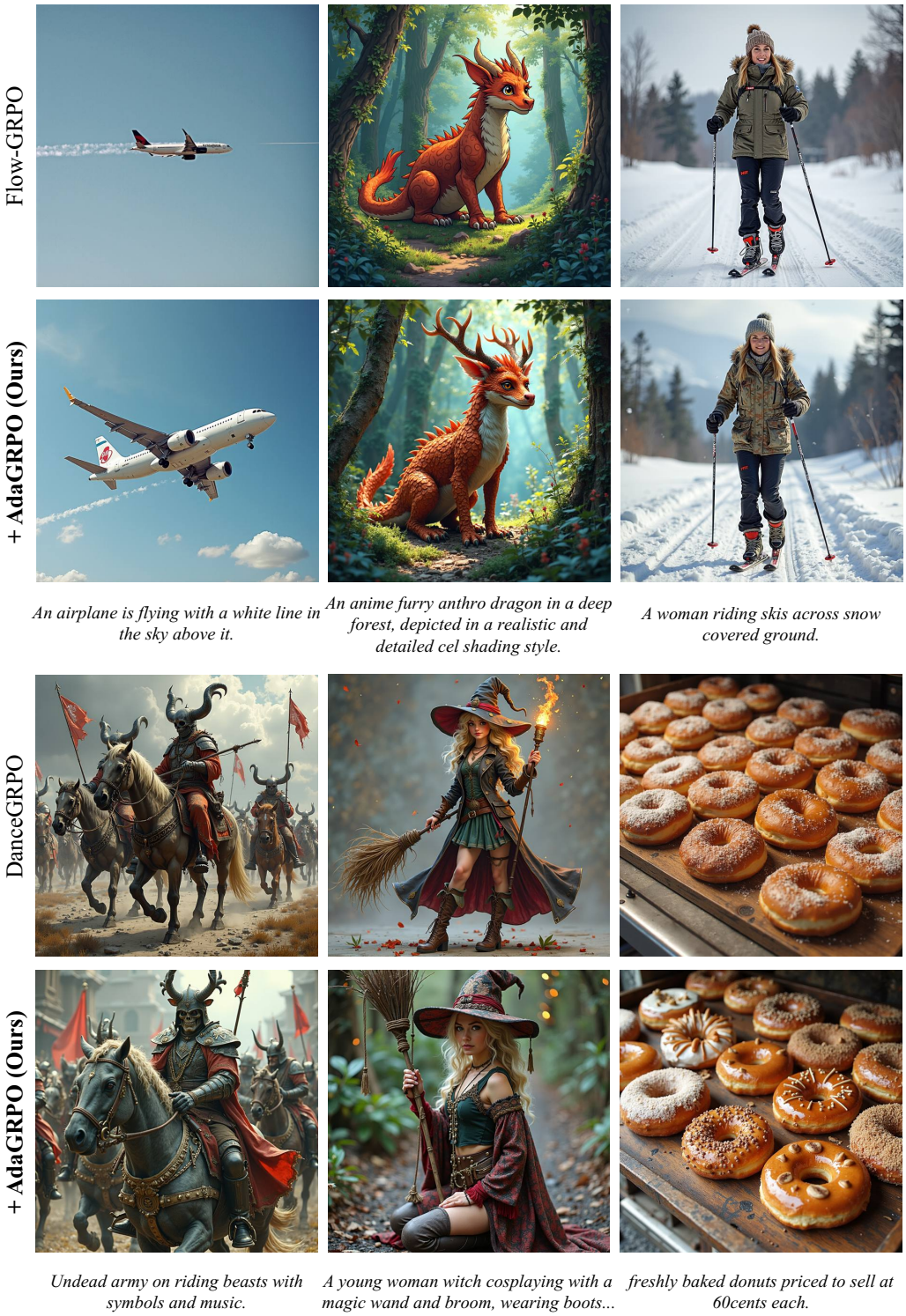


Figure 10: **Additional Comparison Results on HPS-v3. (1/2)**



Figure 11: **Additional Comparison Results on HPS-v3. (2/2)**



Figure 12: **Additional Visual Samples of AdaGRPO. (1/4)**



Figure 13: **Additional Visual Samples of AdaGRPO. (2/4)**



Figure 14: **Additional Visual Samples of AdaGRPO. (3/4)**



Figure 15: **Additional Visual Samples of AdaGRPO. (4/4)**



Figure 16: **Results using same prompts and different seeds. (HPS-v2)**

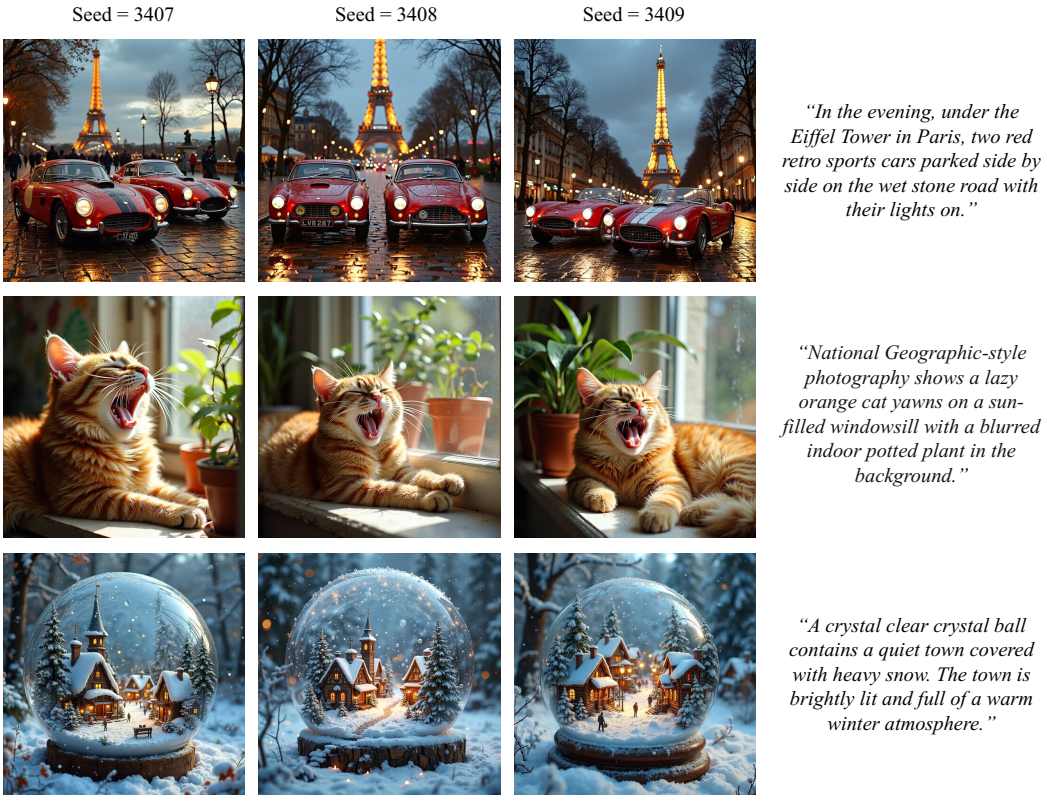


Figure 17: **Results using same prompts and different seeds. (HPS-v3)**

Table 8: The image generation prompts for each figure are listed sequentially, following the order from left to right and top to bottom. (Table 1/3)

Figure	Text Prompt
Figure. 1	There was an open retro wooden jewelry box with an exquisite sapphire necklace lying quietly inside, shining with a glimmer.
	Very ornate bedroom with a chandelier over the bed.
	A spaceship in an empty landscape.
	A Japanese castle landscape painting trending on Artstation.
	A giant panda dressed in the costume of an ancient Egyptian pharaoh sits upright on a golden throne in the dense bamboo forest.
	A photorealistic, ultra-detailed macro shot of an exquisite artisanal chocolate cake resting on an elegant vintage silver cake stand. The glossy, rich chocolate ganache reflects soft, warm ambient lighting. In the center of the cake, the text "Adaptive GRPO" is flawlessly written in elegant white vanilla cream piping. The cake is intricately decorated with delicate edible gold leaf flakes, fresh plump raspberries, and dark chocolate curls. The background is a sophisticated, dimly lit fine dining restaurant with beautiful warm ambient bokeh blurring out crystal glasses and glowing candles. Cinematic lighting, highly appetizing, high-end food photography, 8k resolution.
	A Monet portrait.
	A turtle with feathered wings is not in the air because it cannot fly, but lies leisurely on the clouds to rest.
	On the street of the future of cyberpunk-style Tokyo, a woman wearing VR glasses controls the holographic koi floating in front of her through the air.
	A crystal wall clock in the shape of an ancient Roman Colosseum.
	Please generate a documentary black and white photo of a tiny astronaut standing next to the ancient ruins of a giant planet.
Figure. 2	A bear walks through a group of bushes with a plant in its mouth.
	A hand-drawn cute gnome holding a pumpkin in an autumn disguise, portrayed in a detailed close-up of the face with warm lighting and high detail.
Figure. 5	People standing in the grass playing with a frisbee.
	Two young ladies seated with several other people at a dinner table.
	Hooded figure standing over a ruined city with red haze and a grin.
	Portrait of an anime princess in white and golden clothes.
	Side of a street, where there is a fire hydrant and a mirror showing the street.
	Side-view blue-ice sneaker inspired by Spiderman created by Weta FX.
Figure. 6	A cute anime schoolgirl with a sad face submerged in dark pink and blue water, portrayed in an oil painting style.
	A warrior in glowing azure plate armor stands in a doorway to hell sliced by iridescent glass cracks, with crimson clouds and an art deco palace backdrop.
	A man with his finger held up to his nose.
	A photorealistic image from a furry fandom convention set in a biopunk era after the genetic revolution and quantum singularity.
	A phantom airship.
	A portrait painting of Leighann Vail.
Figure. 7	A cat with two horns on its head.
	A massive and brightly colored spacecraft in a deserted landscape, depicted in retro 1960s sci-fi art.
	A portrait of an anime mecha robot with a Japanese town background and a starred night sky.
	A wooden trunk sitting outside with stickers on it.
	The image is a highly detailed concept art of a medieval city, with a cinematic style and painted beautifully in oil by various artists including Wlop, Greg Rutkowski, and Artgerm, and can be viewed on Artstation.
Figure. 8	A 3D rendering of anime schoolgirls with a sad expression underwater, surrounded by dramatic lighting.
	A couple of men are standing outside their car watching sheep cross a road.
	The image is of Pixel Art Huggy Wuggy performing a jumpscare.
	A detailed and realistic fantasy Proto-Slavic skinny red troll creature.
	A man walking on the bea with his surfboard.
	A fruit basket on a kitchen table with a Studio Ghibli reference.
Figure. 9	A Wojak looking over a sea of memes from a cliff on 4chan.
	A steampunk pocketwatch owl is trapped inside a glass jar buried in sand, surrounded by an hourglass and swirling mist.
	American cowboy with a scruffy appearance in a retrofuturistic style, inspired by the animations of Studio Ghibli.
	A person flying a kite while standing in the grass.
	An image of an emo with dark brown hair in a messy pixie cut, large entirely-black eyes, wearing black clothing and boots.
	Album art of a hand holding a balloon emerging from the water against a red sky.

Table 9: The image generation prompts for each figure are listed sequentially, following the order from left to right and top to bottom. (Table 2/3)

Figure	Text Prompt
Figure. 10	An airplane is flying with a white line in the sky above it.
	An anime furry anthro dragon in a deep forest, depicted in a realistic and detailed cel shading style.
	A woman riding skis across snow covered ground.
	Undead army on riding beasts with symbols and music.
	A young woman witch cosplaying with a magic wand and broom, wearing boots, and posing in a full body shot with a detailed face.
	freshly baked donuts priced to sell at 60cents each.
Figure. 11	Small carved figurines of fantasy buildings, miniatures and standalone objects on a table.
	An image of a corn elemental, created as concept art for a high fantasy setting.
	Fullbody portrait of half-mouse anime girl by A-1 Pictures, trending on ArtStation.
	The Magician by René Magritte.
	The image features baroque architecture by Escher and Jean Delville in the walled city of Kowloon, lit with golden lighting and displaying ornate details.
	Female human barbarian in a dungeon, illustrated by Jeff Easley for Dungeons and Dragons.
Figure. 12	From the interior perspective of the ancient Roman Colosseum, sunlight shines in through the dilapidated roof structure, illuminating the sand and passages below.
	A miniature snowy mountain wonderland enclosed in glass.
	Please design a playback scene for a music App. A cartoon bear wearing a spacesuit floats to touch the musical symbols made of crystals. The overall style is a glass imitation.
	Small carved figurines of fantasy buildings, miniatures and standalone objects on a table.
	A portrait of Frank Zappa smoking, with vivid neon colors, by various artists.
	In a clay sculpture, a hungry fox looks up at an empty bird's nest on a high branch, with broken eggshells scattered on the ground below.
	a man walking alone down the street in a velvet jacket.
	A plate topped with lots of different kinds of fruit.
	In the golden desert at dusk, a traveler carrying a huge backpack is struggling to walk into the distance, trailing a long shadow behind him.
	An ornate box with a heart-shaped jewel inside.
	Please generate a cartoon-style picture of a squirrel holding a heart-shaped nut in his hand, with a confused expression, but it does not open its mouth to eat it.
	An old-fashioned radio was placed on the dusty desk. It didn't play any sound. The diary next to it was open, and the atmosphere was suspenseful.
Figure. 13	A dragon standing in a forest, drinking river water.
	A steampunk picture of a lady in a classic dress operating a complex brass mechanism.
	Vincent Van Gogh stood in the golden wheat field. He looked up at the stars with melancholy and fanaticism on his face, with a cinematic texture like an oil painting.
	A celadon teapot around with colorful butterflies among the hazy distant mountains. It has a strong national style and ink painting style.
	A ginger haired mouse mechanic in blue overalls in a cyberpunk scene with neon slums in the background.
	A photo of a mechanical angel woman with crystal wings, in the sci-fi style of Stefan Kostic, created by Stanley Lau and Artgerm.
	A curious dog greets a colorful parrot leaning out of its wooden cage.
	A massive and brightly colored spacecraft in a deserted landscape, depicted in retro 1960s sci-fi art.
	Generating picture: A steampunk-style robot is stretching out its brass fingers and carefully touching a fragile dandelion.
	A great room with the living area in the foreground, dining table behind it and kitchen in the very back.
	There are two hourglasses on the bookshelf.
	A geometric pink fox resting among blooming flowers.
Figure. 14	In Pixar animation style, a young detective holding a magnifying glass squatted on the ground and carefully observed a pool of spilled milk.
	a castle is in the middle of a eurpean city.
	a papaya fruit dressed as a sailor.
	Under the Eiffel Tower in Paris at night, a quiet street is arranged as a product launch scene. The ground is wet and reflects light, and the overall futuristic technological style is adopted.
	A key visual of a young female swat officer with a neon futuristic gas mask in a cyberpunk setting.
	a plate that has some kind of food on it.
	A lighted birthday cake with chunks of walnuts.
	In the ruined city, a little boy looked up in surprise at the huge, still robot in front of him, with a cinematic lens and a realistic style.
	A bronze steampunk mechanical heart, intertwined with gears and pipes, exuding a faint orange glow, surrealist oil painting.
	An elderly historian wearing white cotton gloves carefully examined a yellowed sheepskin scroll map with a magnifying glass, with a solemn expression.
	A large dog laying on a couch in a room.
	A huge glass greenhouse shaped like the Great Pyramid of Giza contains a complete and miniature Amazon rainforest ecosystem, and the overall surrealist style.

Table 10: The image generation prompts for each figure are listed sequentially, following the order from left to right and top to bottom. (Table 3/3)

Figure	Text Prompt
Figure. 15	Generating picture: depicts the famous scene of physicist Newton sitting under an apple tree, falling into thought after being hit by a falling apple.
	An oil painting portrait of a beautiful dryad wearing an ombre velvet gown, with long hair and a tiara, adorned with dozens of jeweled necklaces, and illuminated with dramatic cinematic lighting.
	An eagle carved out of crystal has huge, colorful maple leaves that spread its wings on the base and want to fly.
	Oil painting portrait of demon king with gazing eyes, art by John Howe, Keith Parkinson, and Larry Elmore, featured on ArtStation and CGSociety.
	Generated image: Inside an ancient church, the dome is composed of countless stained glass sheets, and the picture presents a solemn cinematic realistic style.
	Close-up view of ancient Greek ruins set against a colourful, starry night sky creating a mystical atmosphere.
	A lion lies quietly on the bottom of the sea.
	A train that is going by a building.
	A regal robot sits thoughtfully on the throne.
	A white Persian cat wearing a peacock feather headdress and surrounded by flowers, in a magical realism painting.
	Please draw a castle made of marshmallows, with melted chocolate on top of its towers, and the overall sweet and dreamy style.
	An anime girl is riding a bicycle in Akihabara, resembling the style seen in Studio Ghibli films, and the depiction is detailed.
Figure. 16	Generate a game concept setting diagram: a huge turtle carries a small castle on its back, which serves as a mobile base for players and travels through the fantasy forest.
	An astronaut stood under the glass dome of the huge lunar base, holding a miniature, blue-emitting model of the earth in his hand, with a very focused expression.
	Please generate: In an empty ancient library late at night, moonlight shines down from the high windows, illuminating rows of tall bookshelves and scattered books, presenting the realistic style of suspense movies.
Figure. 17	In the evening, under the Eiffel Tower in Paris, two red retro sports cars parked side by side on the wet stone road with their lights on.
	National Geographic-style photography shows a lazy orange cat yawns on a sun-filled windowsill with a blurred indoor potted plant in the background.
	A crystal clear crystal ball contains a quiet town covered with heavy snow. The town is brightly lit and full of a warm winter atmosphere.