λ-CONTROLLED GRPO: TURNING FLOW-MATCHING RATIO INSTABILITY INTO A BUDGETED RESOURCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning is increasingly used to align image generators with reward signals, and Flow-GRPO recently extended this paradigm to flow-matching models by treating the denoising sampler as a stochastic policy that can be optimized from reward feedback. Training in this setting is unstable in a way specific to multi-step denoising: the policy update changes systematically across denoising steps, with importance ratios drifting below one, becoming increasingly dispersed, clipping at different rates, and leaving fewer usable samples late in training. Prior work treats these effects as separate failure modes and addresses each with a hand-tuned stabilizer. We show instead that they arise from a single per-step quantity, which we call *path variance*. This quantity is determined exactly by the sampler's Gaussian transition kernel and can be estimated cheaply during training. This reframes instability as a resource that can be measured and budgeted rather than a collection of symptoms to repair. Our method, λ-Controlled GRPO, calibrates importance-ratio behavior from this predicted law rather than from noisy empirical statistics, and allocates gradient effort across denoising steps according to their predicted cost. The two scales governing the update are fixed by standard policy choices rather than introduced as free tuning parameters. On a text-to-image model under two reward settings, rendering difficult target text scored by optical character recognition and matching human preferences scored by a preference model, λ-Controlled GRPO improves both text accuracy and preference reward over the strongest empirical stabilizer. It also keeps late-step path variance within its intended budget, precisely where the baseline systematically overshoots. The result is a Flow-GRPO update calibrated by its own transition law rather than stabilized after instability appears.

1 INTRODUCTION

Reinforcement learning has become a standard tool for aligning generative models with reward signals, and has recently extended from language models to the flow-matching and rectified-flow models used for image generation. The key step, Flow-GRPO, converts a deterministic flow-matching sampler into a stochastic policy whose denoising trajectory has a tractable per-step likelihood, enabling policy-gradient updates from the group relative policy optimization (GRPO) family. This makes online reward optimization possible for image generators, but introduces a stability problem specific to multi-step denoising: the policy importance ratio, which measures how much more or less likely the new policy is to take a sampled step than the old one, behaves very differently across denoising timesteps. Recent work reports that ratios shift below one at some steps, become highly dispersed at others, clip asymmetrically in the surrogate objective, and eventually leave few effective samples late in training. These effects have largely been treated as separate empirical symptoms and addressed with hand-designed stabilizers, such as timestep-wise normalization of observed ratios and manually chosen gradient reweighting. What is missing is a common explanation for why these pathologies emerge and which quantity an update should control. We identify such a quantity directly from the finite-grid Gaussian transition kernel used by the sampler. A single per-step scalar, which we call the *path variance*,

$$\lambda_k = \left\| \sigma_k^{-1} (b_\theta - b_{\text{old}})(X_k, t_k) \right\|^2 \Delta t_k,$$

determines the conditional mean and variance of the step log-ratio, and therefore predicts its drift, dispersion, left shift, clipping imbalance, and loss of usable samples within one law. The same

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.

scalar is not only diagnostic. It defines the natural cost of a policy update at each denoising step and therefore provides a principled quantity for allocating gradient effort across the trajectory.

A practical subtlety arises because real implementations typically store a mean-reduced log-probability rather than the full coordinate sum. This reduction splits the path-variance effect into two related quantities: a centering term that governs the mean shift of the log-ratio and a variance term that governs its spread. Keeping these roles separate is essential for the theoretical prediction to match the statistics measured during training.

This paper makes four contributions. First, we derive the finite-grid importance-ratio law and its mean-reduced corollary, obtaining an exact per-step prediction for the log-ratio statistics from the path variance. Second, we give a cheap online estimator of this quantity using values already produced by the sampler. Third, we turn the prediction into an algorithm, λ-Controlled GRPO, which replaces empirical ratio normalization with analytic calibration and allocates gradient effort according to predicted path-variance cost, with both governing scales determined by standard policy choices rather than introduced as free hyperparameters. Fourth, across two reward regimes for a text-to-image model, we show that the analytic update improves task quality over the strongest empirical stabilizer while keeping late-step path-variance spend near its intended budget. Throughout, we first verify that the estimated scalar predicts the measured log-ratio statistics before modifying the optimizer, so the resulting intervention is grounded in a confirmed transition-law prediction rather than an assumed training heuristic.

2 Related work

Reinforcement learning for flow-matching models. Reinforcement learning has become a standard tool for aligning generative models with rewards, beginning in language with reward models learned from human preferences (Christiano et al., 2017; Ouyang et al., 2022) and policy or preference optimization methods including trust-region and proximal policy optimization (Schulman et al., 2015; 2017), group relative policy optimization (GRPO) (Shao et al., 2024; Guo et al., 2025), and direct preference optimization (Rafailov et al., 2023). The same principle was extended to image diffusion models by treating denoising as a sequential decision process optimized with policy gradients (Black et al., 2024; Fan et al., 2023), differentiating through the sampler under differentiable rewards (Clark et al., 2024; Prabhudesai et al., 2023; Eyring et al., 2024), and adapting preference optimization to diffusion models (Wallace et al., 2024). Flow-GRPO (Liu et al., 2025) and DanceGRPO (Xue et al., 2025) bring GRPO to flow-matching image generators by converting the deterministic sampler into a stochastic policy through an ODE-to-SDE construction with Gaussian transition likelihoods and online updates, making sampled denoising trajectories admit tractable per-step importance ratios. This is the policy class studied here, and it raises a central question: how much policy movement can each denoising step absorb before its likelihood ratio becomes unstable? The closest empirical antecedent is GRPO-Guard (Wang et al., 2025a), which reports the same ratio pathologies studied here, including implicit over-optimization, ratios shifted below one, timestep inconsistency, and clipping imbalance, and stabilizes training through empirical ratio normalization and regulated clipping. That work establishes the failure mode, but its correction is not derived from the transition kernel. It recenters measured ratios after instability appears rather than identifying a quantity that predicts their mean shift and variance before the update, or determines how much gradient a timestep should receive per unit cost. Our finite-grid analysis (Section 3.2) supplies that quantity: conditioned on the current state, the step log-ratio is Gaussian with mean $-\lambda_k/2$ and variance λ_k. A single λ_k therefore determines the typical left shift, ratio dispersion, timestep inconsistency, and clipping imbalance together. Empirical normalization and gradient reweighting can thus be viewed as symptom-level corrections to a quantity that is predictable from the transition law itself.

Timestep, sampling, and surrogate-ratio variants. A rapidly growing literature improves Flow-GRPO from complementary directions. Several methods modify where, when, or how denoising trajectories are sampled or weighted. TempFlow-GRPO (He et al., 2025) emphasizes timestep allocation, DenseGRPO (Deng et al., 2026) densifies reward feedback across denoising steps, Smart-GRPO (Yu et al., 2025) searches over noise perturbations, BranchGRPO (Li et al., 2025b) introduces structured branching, MixGRPO (Li et al., 2025a) changes the ODE-to-SDE mixture, CPS (Wang & Yu, 2025) studies coefficient-preserving sampling, Pref-GRPO (Wang et al., 2025b) replaces point-

wise rewards with pairwise preferences to improve optimization stability, OP-GRPO (Zhang et al., 2026) targets replay and off-policy efficiency, and Flow-Factory (Ping et al., 2026) organizes the broader design space. These works demonstrate that sampling and credit assignment strongly affect optimization, but none derives a per-step likelihood-ratio resource from the Gaussian transition kernel and uses that quantity to calibrate both ratio statistics and gradient allocation. A λ budget is therefore largely complementary to these approaches. FPO (McAllister et al., 2025) takes a different route by replacing exact likelihood computation with a flow-matching surrogate ratio, which is useful when exact transition likelihoods are unavailable. In Flow-GRPO, however, the Euler–Maruyama Gaussian transition is explicit, allowing us to analyze the exact finite-grid ratio and use its path-variance scalar as a predictive optimization budget. Prior work therefore provides the policy class, empirical stabilizers, sampling strategies, and surrogate objectives; our contribution is to connect these phenomena through a single analytically tractable quantity that both predicts per-step ratio behavior and determines how gradient effort should be allocated.

3 Preliminaries and Methods

3.1 Preliminaries

Flow matching. A rectified-flow model (Liu et al., 2023; Lipman et al., 2023; Albergo et al., 2025) transports data to noise along the linear interpolant

$$x_t = (1-t)x_0 + tx_1, \qquad x_0 \sim p_{\text{data}}, \quad x_1 \sim \mathcal{N}(0, I), \quad t \in [0,1], \tag{3.1}$$

and learns a velocity field $v_\theta(x,t)$ that predicts $x_1 - x_0$. Deterministic generation then solves the ordinary differential equation $\mathrm{d}X_t = v_\theta(X_t,t)\,\mathrm{d}t$. Rectified flow and closely related diffusion models (Ho et al., 2020; Song et al., 2021) underlie modern text-to-image generators such as SD3 (Esser et al., 2024) and FLUX (Labs et al., 2025), whose transformer backbones follow the diffusion-transformer line (Peebles & Xie, 2023; Ma et al., 2024).

Stochastic policy. GRPO requires stochastic trajectories with tractable transition densities. Flow-GRPO (Liu et al., 2025) therefore replaces the deterministic ODE with a marginal-preserving stochastic differential equation, $\mathrm{d}X_t = b_\theta(X_t,t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t$, whose drift b_θ is fixed through Fokker–Planck matching to the flow-matching score and whose diffusion σ_t follows a prescribed schedule, both derived in Section C. Training discretizes this process on a grid $0 = t_0 < \cdots < t_K = 1$ with step sizes Δt_k, yielding the Euler–Maruyama Gaussian transition kernel

$$K_\theta(x_{k+1} \mid x_k) = \mathcal{N}\big(x_k + b_\theta(x_k,t_k)\Delta t_k,\ \sigma_k\sigma_k^\top \Delta t_k\big), \tag{3.2}$$

which is the central object of our analysis.

GRPO objective. For a prompt c, a group of G trajectories is sampled from the old policy and evaluated by a reward model, producing group-relative advantages $\widehat{A}^i$ (Shao et al., 2024), with rewards standardized within each group. Each sampled transition $x_k^i \to x_{k+1}^i$ carries the per-step importance ratio

$$r_k^i(\theta) = \frac{K_\theta(x_{k+1}^i \mid x_k^i, c)}{K_{\text{old}}(x_{k+1}^i \mid x_k^i, c)}, \tag{3.3}$$

and the clipped surrogate (Schulman et al., 2017) optimizes $\min\big(r_k^i\widehat{A}^i,\ \text{clip}(r_k^i, 1-\epsilon, 1+\epsilon)\widehat{A}^i\big)$ with clipping radius ϵ. The next subsection characterizes how r_k varies across denoising timesteps.

3.2 The finite-grid path-ratio law

The behavior of r_k is determined exactly by the transition kernel under one assumption satisfied by the fixed-diffusion schedules used in Flow-GRPO.

Assumption 3.1 (Fixed diffusion during the policy update)**.** At each timestep, the new and old transition kernels share the same diffusion $\sigma_k\sigma_k^\top$ and differ only in their drift.

A direct Gaussian calculation (Section D) then yields the one-step log-ratio in closed form.

Proposition 3.2 (One-step Gaussian ratio identity)**.** *Let* $h_k = \sigma_k^{-1}(b_\theta - b_{\text{old}})(X_k,t_k)$ *and* $\Delta W_k \sim \mathcal{N}(0, \Delta t_k I)$*. Then*

$$\xi_k := \log \frac{K_\theta(X_{k+1} \mid X_k)}{K_{\text{old}}(X_{k+1} \mid X_k)} = \langle h_k, \Delta W_k\rangle - \tfrac{1}{2}\,\|h_k\|^2\,\Delta t_k. \tag{3.4}$$

The squared coefficient defines the quantity that drives the paper, the *per-step path variance*

$$\lambda_k \coloneqq \|h_k\|^2 \Delta t_k = \left\|\sigma_k^{-1}(b_\theta - b_{\text{old}})(X_k, t_k)\right\|^2 \Delta t_k, \tag{3.5}$$

which measures the squared policy displacement at timestep k in units of the transition noise, equivalently the local Mahalanobis distance between the new and old drifts. This single scalar determines the complete conditional ratio law.

Proposition 3.3 (Conditional lognormal ratio law)**.** *Conditioned on $\mathcal{F}_{t_k}$, the one-step log-ratio is Gaussian,*

$$\xi_k \mid \mathcal{F}_{t_k} \sim \mathcal{N}\big(-\tfrac{\lambda_k}{2}, \lambda_k\big)\,, \tag{3.6}$$

so, with $r_k = \exp(\xi_k)$,

$$\mathbb{E}[r_k \mid \mathcal{F}_{t_k}] = 1, \quad \text{median}(r_k \mid \mathcal{F}_{t_k}) = e^{-\lambda_k/2}, \quad \text{Var}(r_k \mid \mathcal{F}_{t_k}) = e^{\lambda_k} - 1. \tag{3.7}$$

A single λ_k therefore controls the negative log-ratio drift, the left-shifted typical ratio, the ratio variance, the imbalance between the two PPO clipping tails, and the loss of effective samples. Closed-form clipping-tail expressions are given in Section D. Empirical stabilizers from prior work can therefore be viewed as correcting several symptoms of one analytically predictable quantity.

One implementation detail is essential. Practical Flow-GRPO code stores a mean-reduced log-probability rather than the full coordinate sum. Let $a_{k,m} = (\mu_{\theta,k,m} - \mu_{\text{old},k,m})/s_{k,m}$ denote the normalized mean shift for coordinate m. The mean-reduced log-ratio is

$$\bar{\xi}_k = \frac{1}{D}\sum_{m=1}^{D}\Big(a_{k,m}\epsilon_{k,m} - \tfrac{1}{2}a_{k,m}^2\Big), \qquad \epsilon_{k,m} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{3.8}$$

and the reduction separates the path-variance effect into a centering scale and a variance scale,

$$\lambda_k^{\text{center}} = \frac{1}{D}\sum_m a_{k,m}^2, \qquad \lambda_k^{\text{var}} = \frac{1}{D^2}\sum_m a_{k,m}^2. \tag{3.9}$$

Proposition 3.4 (Reduced-log-prob ratio law)**.** *For the mean-reduced log-ratio in Equation (3.8), conditioned on $\mathcal{F}_{t_k}$, $\mathbb{E}[\bar{\xi}_k] = -\frac{1}{2}\lambda_k^{\text{center}}$, $\text{Var}(\bar{\xi}_k) = \lambda_k^{\text{var}}$.*

The two scales differ by the latent dimension. This explains why the predicted mean-shift law remains pronounced while the raw variance curve can appear nearly flat, and why centering and variance must be treated separately in the optimizer.

3.3 λ-Controlled GRPO

Our method estimates the path variance online and treats it as a resource to budget across denoising timesteps. It has two analytic components, and the constants governing both are determined by standard policy choices rather than tuned as free hyperparameters. Full derivations are given in Sections C and E.

Estimating the scalar. The transition means $\mu_{\theta,k} = X_k + b_\theta \Delta t_k$ and $\mu_{\text{old},k}$ are already available during sampling, so the reduction-matched estimators add negligible cost:

$$\widehat{\lambda}_{i,k}^{\text{center}} = \frac{1}{D}\sum_{m=1}^{D}\Big(\frac{\mu_{\theta,k,m}^i - \mu_{\text{old},k,m}^i}{s_{k,m}}\Big)^2, \qquad \widehat{\lambda}_{i,k}^{\text{var}} = \frac{1}{D^2}\sum_{m=1}^{D}\Big(\frac{\mu_{\theta,k,m}^i - \mu_{\text{old},k,m}^i}{s_{k,m}}\Big)^2, \tag{3.10}$$

where $s_{k,m} = \sigma_{k,m}\sqrt{\Delta t_k}$. Section C gives the endpoint scaling of λ_k under the Flow-GRPO schedule.

LambdaNorm-T: analytic ratio calibration. Because exponentiating the mean-reduced log-ratio does not preserve the mean-one property of the full likelihood ratio, we calibrate it from its predicted law instead of reusing it directly. We center and standardize $y_{i,k} = \bar{\xi}_{i,k}$ using the predicted moments and then rescale to a target variance ν:

$$z_{i,k} = \frac{y_{i,k} + \frac{1}{2}\widehat{\lambda}_{i,k}^{\text{center}}}{\sqrt{\widehat{\lambda}_{i,k}^{\text{var}} + \varepsilon_\lambda}}, \qquad \widetilde{r}_{i,k} = \exp\big(\sqrt{\nu}\, z_{i,k} - \tfrac{1}{2}\nu\big). \tag{3.11}$$

If $z_{i,k}$ is standard normal, then $\mathbb{E}[\widetilde{r}_{i,k}] \approx 1$, so $\widetilde{r}_{i,k}$ becomes a calibrated, mean-one surrogate ratio. The scale ν is not introduced as an independent tuning parameter. PPO already specifies a clipping radius ϵ, so we set $\nu = \epsilon^2$.

Damp-only λ-gradient weighting. We next allocate gradient effort according to predicted path-variance cost. A timestep is damped only when its predicted center exceeds a budget τ:

$$w_k^\lambda = \min\Big(1,\ \sqrt{\tau/(\widehat{\lambda}_k^{\text{center}} + \varepsilon_\lambda)}\Big), \qquad \mathcal{L} = -\,\mathbb{E}_{i,k}\Big[w_k^\lambda \min\Big(\widetilde{r}_{i,k}\widehat{A}_i,\ \text{clip}(\widetilde{r}_{i,k}, 1-\epsilon, 1+\epsilon)\widehat{A}_i\Big)\Big], \tag{3.12}$$

where $\widehat{\lambda}_k^{\text{center}}$ is the batch mean of the per-sample estimates in Equation (3.10), and all estimates are detached so timesteps already within budget are never amplified. The budget is determined by a target retained effective-sample fraction q (Elvira et al., 2018) over the late timesteps $\mathcal{K}_{\text{late}}$. Under the lognormal path-weight heuristic, $\tau = -\frac{\log q}{|\mathcal{K}_{\text{late}}|}$ (Section E). We report realized spend using the mean late-step path variance

$$\Lambda_{\text{late}} = \frac{1}{|\mathcal{K}_{\text{late}}|} \sum_{k \in \mathcal{K}_{\text{late}}} \widehat{\lambda}_k^{\text{center}}, \tag{3.13}$$

which the budget τ is intended to control. The complete intervention is therefore specified by two interpretable policy choices, the PPO clipping radius ϵ and the retained fraction q. We use $\epsilon = 0.2$ and $q = 0.95$ for OCR, and the tighter setting $\epsilon = 0.1$ and $q = 0.99$ for the denser PickScore reward.

4 Experimental findings

The experiments address three questions. tiny-SD3 diagnostics test whether the finite-grid law predicts training log-ratio statistics. Primal-dual experiments test whether path variance is controllable, not just descriptive. SD3.5 experiments test whether the analytic method, LambdaNorm-T with damp-only λ-gradient weighting, improves task quality over the empirical stabilizer under matched budgets. The main held-out comparison appears in Table 2.

4.1 Tiny-SD3 law audit

The tiny-SD3 audit tests whether the finite-grid path-variance law appears in a working Flow-GRPO implementation. We train a small public SD3 debug pipeline with a compressibility reward and average every statistic over three independent runs from fresh random initializations (configuration in Section B). Here, $\log r_k$ denotes the mean-reduced log-ratio $\bar{\xi}_k$ from Equation (3.8), which is the quantity stored by the implementation. In the baseline run, the drift-based estimator $\widehat{\lambda}_k^{\text{center}}$ closely tracks the negative log-ratio drift, as predicted by the reduced law $\widehat{\lambda}_k^{\text{center}} \approx -2\,\widehat{\mathbb{E}}[\log r_k]$. At the final timestep, for example, $\widehat{\lambda}_7^{\text{center}} = 0.0054$ predicts a drift of -0.0027, exactly matching the measured -0.0027. The corresponding $\widehat{\lambda}_k^{\text{var}}$ also lies on the correct scale for the observed variance (Figure 1). This agreement provides the first direct evidence that the derived law predicts statistics measurable during training.

We next add λ-z clipping to test whether controlling the predicted quantity also controls the measured pathology. Averaged over three seeds and the final 50 updates, it reduces the late-step $\widehat{\lambda}_7^{\text{center}}$ from 0.0054 to 1.0×10^{-4} and the log-ratio drift from -0.0027 to -4×10^{-5}, while leaving reward essentially unchanged (Figure 2; full statistics in Table 1). Because this experiment uses a small debug model whose effective sample size remains near one, it is a law audit rather than a performance result. Its purpose is to establish that $\widehat{\lambda}_k^{\text{center}}$ and $\widehat{\lambda}_k^{\text{var}}$ act as accurate online predictors of the reduced log-ratio center and variance, and that directly reducing λ suppresses the corresponding pathology.

4.2 Tiny-SD3 stress result: λ is controllable

We next test whether λ is actionable rather than merely descriptive. As a control ablation, we impose a primal-dual λ budget (Section G). Across three paired tiny-SD3 stress seeds, the controller improves the compressibility reward on every seed while reducing the average final-step $\widehat{\lambda}_7^{\text{center}}$ from 0.998 to 0.166. This result establishes that the derived scalar can steer optimization. The main

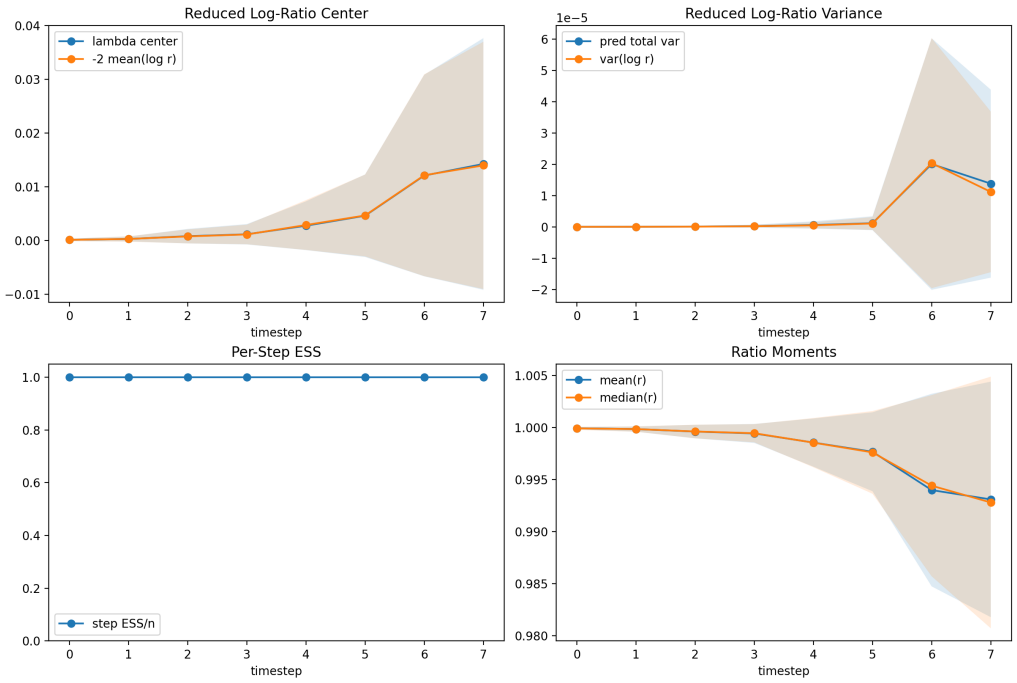


Figure 1: Baseline diagnostics from the tiny-SD3 law audit. The key comparisons are between $\widehat{\lambda}_k^{\text{center}}$ and $-2\widehat{\mathbb{E}}[\log r_k]$, and between $\widehat{\lambda}_k^{\text{var}}$ and $\widehat{\text{Var}}(\log r_k)$ under the reduced-log-prob convention.

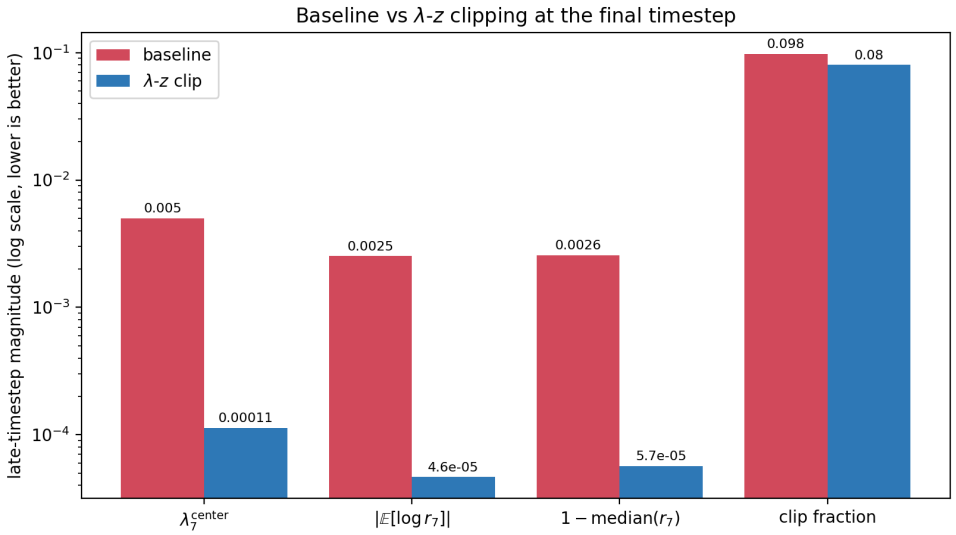


Figure 2: Baseline versus λ-z clipping in the tiny-SD3 law audit. λ-z clipping sharply reduces the late-step $\widehat{\lambda}_k^{\text{center}}$, $\widehat{\lambda}_k^{\text{var}}$, log-ratio mean drift, reduced log-ratio variance, and typical-ratio depression.

Metric at timestep 7	Baseline	λ-z clip
$\widehat{\lambda}_7^{\text{center}}$	0.005425	0.000100
$\widehat{\lambda}_7^{\text{var}}$	3.31×10^{-7}	6.12×10^{-9}
$\widehat{\mathbb{E}}[\log r_7]$	-0.002693	-0.000041
$\widehat{\mathrm{Var}}(\log r_7)$	7.32×10^{-7}	8.83×10^{-9}
Median ratio	0.997264	0.999949
Per-step ESS / n	0.9999993	0.99999996
Path ESS / n	0.999996	1.000000
Clip fraction	0.1169	0.0730
Reward average	-0.008114	-0.008238

Table 1: Tiny-SD3 law audit at the final timestep. Statistics are averaged over three seeds and the final 50 updates. λ-z clipping reduces predicted path variance, log-ratio drift and variance, and typical-ratio depression while leaving reward essentially unchanged.

method below removes the dual machinery and instead uses the analytic prediction directly through LambdaNorm-T and damp-only weighting.

4.3 SD3.5 HARD-OCR RESULTS

Our main real-model experiment fine-tunes SD3.5 Medium with low-rank adaptation (LoRA) on the hard-OCR task. We use the same OCR dataset and reward definition as the public Flow-GRPO and GRPO-Guard configurations, but with smaller update statistics, namely one GPU, smaller groups, fewer batches per epoch, and shorter training. Checkpoints are saved at fixed training-step intervals and selected using a held-out protocol. A random 256 OCR test prompts are used only for checkpoint selection, while the remaining 762 prompts are reserved for final evaluation. Alongside the Flow-GRPO-style OCR reward, we report character error rate (CER), defined as the fraction of characters that must be edited to recover the target, together with exact-match and substring-match rates. The latter require the complete target or a contiguous portion of it to appear in the generated image. PickScore (Kirstain et al., 2023) and CLIP score (Hessel et al., 2021) serve as optional image-quality guardrails. The primal-dual controller from Section G is included only as a baseline, denoted "Dual," alongside empirical RatioNorm.

Under this protocol, empirical RatioNorm selects checkpoint 40, before the late-training pathology emerges, while the analytic method and primal-dual baseline select checkpoint 80. On the held-out 762-prompt test split, LambdaNorm-T with damp-only λ weighting outperforms both baselines on OCR reward, CER, exact match, and substring match (Table 2).

Method	Ckpt.	OCR ↑	CER ↓	Exact ↑	Substr. ↑	Λ_{late} ↓
RatioNorm	40	0.5632	0.4368	12.99%	19.29%	0.0
Dual	80	0.5501	0.4499	12.07%	16.80%	0.000217
λ-Controlled GRPO	80	**0.5831**	**0.4169**	**14.44%**	**22.83%**	0.000332
Δ (λ-Ctrl − RatioNorm)		+0.0199	−0.0199	+1.45	+3.54	
95% bootstrap CI		[+0.006, +0.033]	[−0.061, −0.014]	[+0.3, +6.7]	[+1.7, +8.8]	

Table 2: SD3.5 hard-OCR held-out results on 762 test prompts. The analytic LambdaNorm-T plus damp-only λ-gradient method improves over the early-stopped empirical RatioNorm checkpoint on OCR reward, exact match, and substring match, while keeping realized late-step spend Λ_{late} below the target budget $\tau = -\log(0.95)/5 \approx 0.01026$. The last two rows give the paired difference between λ-Controlled GRPO and RatioNorm (upper row) and its 95% bootstrap confidence interval (lower row; 10,000 resamples over the 762 prompts), with the exact- and substring-match differences expressed in percentage points.

On the 256-prompt validation split used for checkpoint selection, the same method, LambdaNorm-T with damp-only weighting at $\epsilon = 0.2$ and $q = 0.95$, improves the stricter transcript-success metrics much more strongly than the mean OCR reward (Table 3).

Method	Ckpt.	OCR ↑	CER ↓	Exact ↑	Substr. ↑	Λ_{late} ↓
RatioNorm	40	**0.5744**	**0.4256**	10.94%	14.84%	**0.0**
Dual	80	0.5643	0.4357	8.59%	14.06%	0.000217
λ-Controlled GRPO	80	0.5657	0.4343	**15.23%**	**23.83%**	0.000332

Table 3: Random 256-prompt SD3.5 hard-OCR validation comparison. The analytic λ-controlled method substantially improves exact-match and substring-match rates over empirical RatioNorm and the primal-dual ablation while keeping predicted late-step path-variance spend near zero. Relative to RatioNorm, exact match improves by approximately 39% and substring match by approximately 61%.

The SD3.5 hard-OCR experiment provides the strongest evidence for the main algorithmic claim. Replacing empirical RatioNorm with the finite-grid Gaussian calibration of LambdaNorm-T and shaping gradient allocation using the predicted λ cost in Equation (3.12) improves the metrics that directly test whether the requested text appears in the image. On the held-out 762-prompt test split, λ-Controlled GRPO gains $+0.0199$ OCR reward, $+1.44$ exact-match points, and $+3.54$ substring-match points over RatioNorm. On the 256-prompt validation split, the strict-metric gains are larger still, approximately 39% for exact match and 61% for substring match. The law audit already established that $\widehat{\lambda}_k^{\text{center}}$ predicts reduced log-ratio drift. Here, using that same prediction to calibrate ratios and allocate gradient effort translates into improved task performance, while the primal-dual runs serve only as a controllability ablation.

4.4 SD3.5 PickScore results

To test whether the benefit is specific to OCR, we repeat the protocol with PickScore (Kirstain et al., 2023), a human-preference reward model in the same broad family as ImageReward (Xu et al., 2023) and HPSv2 (Wu et al., 2023). Because this reward is denser, we use the tighter operating point $\epsilon = 0.1$ and $q = 0.99$, giving $\nu = 0.01$ and $\tau \approx 2.0 \times 10^{-3}$. These values are fixed before training. We compare against empirical RatioNorm on the identical training pipeline. Every saved checkpoint at steps 40, 60, and 80 is evaluated on a random 256-prompt validation split. Each method's best checkpoint is selected by mean PickScore and then evaluated once on the disjoint 762-prompt test complement.

Method	Ckpt.	Mean PickScore ↑	Std.	Prompt-level comparison
RatioNorm	40	0.8283	0.057	—
λ-Controlled GRPO	80	**0.8358**	0.056	wins 467/762 (61.3%)
Paired Δ (Ours − RatioNorm): $+0.0075$, 95% bootstrap CI $[+0.0059, +0.0092]$, $t \approx 9.1$				

Table 4: Disjoint 762-prompt SD3.5 PickScore test split. Representative checkpoints are selected using Table 6 and evaluated once on the held-out complement. The paired 95% bootstrap confidence interval excludes zero, and λ-Controlled GRPO beats RatioNorm on 61.3% of prompts with no ties. RatioNorm selects checkpoint 40 on validation; its later checkpoints perform worse and are therefore not evaluated on the test split.

The PickScore experiment reproduces the same optimization mechanism observed in OCR. Empirical RatioNorm has no explicit path-variance budget, so its late-step $\widehat{\lambda}^{\text{center}}$ rises silently to approximately 3.4τ by step 80, while mean PickScore falls from 0.8325 at step 40 to 0.8126 at step 80. By contrast, λ-Controlled GRPO keeps late-step path variance near τ and improves monotonically over training. On the held-out test split, the paired bootstrap separates the $+0.0075$ gain from zero across 762 prompts, and the analytic method wins on a clear majority of prompt-level comparisons.

4.5 Generalization to a second backbone: FLUX

To test whether the analytic method transfers to another flow-matching backbone, we additionally use a broader campaign which trains both SD3.5 and FLUX.1-dev on three tasks, PickScore, hard-OCR, and the GenEval compositional-generation benchmark (Ghosh et al., 2023), and compares

λ-Controlled GRPO against empirical RatioNorm and an additional untuned Flow-GRPO baseline. The held-out FLUX results are summarized in Table 5, and matched held-out GenEval image comparisons for the SD3.5 runs of this campaign are shown in Section J.

	Primary metric ↑			Training clip fraction ↓		
FLUX task	Flow-GRPO	RatioNorm	λ-Ctrl	Flow-GRPO	RatioNorm	λ-Ctrl
PickScore	**0.8430**	0.8383	0.8429	0.961	0.888	**0.089**
OCR reward	0.6434	0.6453	**0.6606**	0.965	0.964	**0.060**
GenEval	0.6545	**0.6569**	0.6538	0.990	0.915	**0.100**

Table 5: Held-out FLUX.1-dev results from the broader single-seed 8-GPU A6000 reproduction campaign. We compare λ-Controlled GRPO with untuned Flow-GRPO and empirical RatioNorm on PickScore (762 paired prompts), hard-OCR (762 prompts), and GenEval (2,148 images). The primary metric is mean PickScore, mean OCR reward, or mean GenEval composite score. Training clip fraction is averaged over the 50 updates preceding the selected checkpoint. Paired differences of λ-Controlled GRPO with 95% bootstrap confidence intervals (10,000 resamples; an interval excluding zero indicates a resolved difference) are, versus RatioNorm, $+0.0046\,[+0.0035, +0.0056]$ (PickScore), $+0.015\,[-0.011, +0.042]$ (OCR reward), and $-0.003\,[-0.023, +0.017]$ (GenEval); and versus Flow-GRPO, $-0.0001\,[-0.0013, +0.0010]$, $+0.017\,[-0.010, +0.044]$, and $-0.001\,[-0.024, +0.023]$, respectively. The analytic method improves over RatioNorm on PickScore and OCR and reduces clipping by approximately an order of magnitude on all three tasks; on GenEval, the primary metrics are statistically indistinguishable.

The FLUX results carry the SD3.5 observations over to a second backbone, though the gains are not consistent across tasks. On FLUX PickScore, λ-Controlled GRPO is effectively on par with the untuned Flow-GRPO baseline (difference -0.0001, 95% CI $[-0.0013, +0.0011]$), while surpassing empirical RatioNorm by $+0.0046$ (CI $[+0.0036, +0.0057]$) with a 63.4% prompt-level win rate. For FLUX hard-OCR, the analytic approach attains the highest measured reward and strict scores, but under this single-seed setting the confidence intervals include zero, so the improvement is suggestive rather than conclusive. On GenEval, all three approaches remain statistically comparable. The clearest cross-backbone result is mechanistic: across all three FLUX tasks, the analytic method reduces the necessary training clip fraction by roughly an order of magnitude, matching the stabilization behavior previously observed on SD3.5. This quantity is the calibrated surrogate ratio targeted by our optimization, not a raw ratio computed in an identical way across methods. For that reason, we treat FLUX as directional evidence for generalization across backbones—most strongly for the predicted clipping dynamics, rather than as confirmation of uniform task-level dominance.

5 Conclusion and Limitations

We ask whether Flow-GRPO's timestep-dependent ratio instability has a single controllable cause. It does. Using the finite-grid transition kernel, we show that negative log-ratio drift, timestep-varying variance, clipping imbalance, and sample loss are all driven by one per-step scalar: the path variance λ_k, which sets the conditional mean and variance of the step log-ratio. Prior "stabilizers" adjust λ_k indirectly. We instead control λ_k directly by calibrating ratios from the predicted transition law and allocating gradient effort using predicted path-variance cost, with scales set by the PPO clipping radius and a target sample-retention rate (not free tuning). Because the sampler measures path variance, it can be estimated online, budgeted across timesteps, and "spent" via optimization, yielding updates calibrated to the transition law rather than patched empirically.

These claims hold in a specific regime. The theory is cleanest when both policies share the same diffusion coefficient; policy-dependent diffusion, reverse-time conventions, or modified schedules need a matching finite-grid analysis. Estimated path variance may deviate from realized statistics due to approximate scores, discretization, or implementation details, so we empirically validate the predicted law before using it for control. Evidence is limited: SD3.5 results are from a single-GPU pilot, and a broader single-seed study shows the largest gains for human-preference optimization, favorable but inconclusive text-rendering results, and neutral GenEval performance. In this range, measuring and budgeting path variance explains when and why Flow-GRPO updates remain controlled.

Reproducibility statement

The theoretical claims are stated with their assumptions in Section 3.2 and proved in full in Section D. The estimator of the path-variance scalar and the two policy-derived scales that fix the method are specified in Section 3.3, and the exact experimental configuration, including the models, timestep counts, optimization schedules, reward definitions, and the per-regime operating points, is collected in Section B. Code to reproduce the training and evaluation is provided as supplementary material.

Ethics statement

This work studies the optimization stability of reinforcement learning for text-to-image models and does not involve human subjects or the release of new data. The image generators and reward models used are existing public artifacts, and the reward-optimization procedure inherits their known limitations, including the potential to amplify biases present in the underlying models. We see no additional ethical concerns specific to the analysis presented here.

AI use statement

A large language model assisted the authors in writing and polishing the language of the manuscript; the authors verified that all claims, proofs, mathematical formulations, and reported values are valid, checked them against the implementation and results, and take full responsibility for the manuscript.

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research (JMLR)*, 26, 2025. arXiv:2303.08797.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

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. *arXiv preprint arXiv:2601.20218*, 2026.

Víctor Elvira, Luca Martino, and Christian P. Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2018. doi: 10.1111/insr.12500. arXiv:1809.04129.

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 *International Conference on Machine Learning (ICML)*, 2024.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, et al. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.1038/s41586-025-09422-z.

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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. 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.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online 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. In *International Conference on Learning Representations (ICLR)*, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.52202/068431-2011.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. doi: 10.1109/ICCV51070.2023.00387.

Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-Factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-2338.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015.

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

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, et al. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.00786.

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

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, Meng Wang, Pengfei Wan, and Xiaodan Liang. GRPO-Guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 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.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhao, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Benjamin Yu, Jackie Liu, and Justin Cui. Smart-GRPO: Smartly sampling noise for efficient RL of flow-matching models. *arXiv preprint arXiv:2510.02654*, 2025.

Liyu Zhang, Kehan Li, Tingrui Han, Tao Zhao, Yuxuan Sheng, Shibo He, and Chao Li. OP-GRPO: Efficient off-policy GRPO for flow-matching models. *arXiv preprint arXiv:2604.04142*, 2026.

A Overview of the appendices

The appendices are grouped by purpose. The first group covers methodology and theory: the experimental configuration, the flow-matching preliminaries and drift, the derivations and proofs behind the finite-grid law, the full derivation of the analytic ratio calibration, and the diagnostic and alternative estimators of the path-variance scalar. The next section gives the primal-dual baseline and the diagnostic clipping rule used in the law audit, which appear in the main text only as comparisons. The last three are experimental evidence, presenting the qualitative held-out comparisons for the OCR, PickScore, and GenEval reward regimes.

B Experimental configuration

The tiny SD3 law audit was deliberately small. It used a tiny public SD3 debug pipeline with a JPEG-compressibility reward and $K = 8$ denoising timesteps, trained for 50 epochs with 32 samples per epoch, 5 inner epochs, and a batch size of 8. The two matched configurations were a baseline diagnostic run and a λ-z-clipping run, and every reported statistic is the average of the last 50 updates over 3 seeds, where a seed is an independent training run from a fresh random initialization.

The SD3.5 experiments fine-tuned SD3.5 Medium with low-rank adaptation on a single GPU, using the same OCR and PickScore datasets and reward definitions as the public Flow-GRPO and GRPO-Guard configurations, but with smaller groups, fewer batches per epoch, and shorter training than the original multi-GPU regime. The two policy choices that define the method were fixed before training in each reward regime: the PPO step radius ϵ and the effective-sample-size retention target q, from which $\nu = \epsilon^2$ and $\tau = -\log(q)/|\mathcal{K}_{\text{late}}|$ follow. The OCR runs used $\epsilon = 0.2$ and $q = 0.95$; the denser PickScore reward used the tighter $\epsilon = 0.1$ and $q = 0.99$.

C Flow-matching preliminaries and drift

Section 3 states the marginal-preserving SDE and its Gaussian transition kernel. This appendix gives the drift, the diffusion schedule, and the endpoint behavior of the path-variance scalar.

The deterministic flow-matching sampler solves

$$\mathrm{d}X_t = v_\theta(X_t, t)\,\mathrm{d}t. \tag{C.1}$$

GRPO needs stochastic trajectories with a Gaussian transition density, so Flow-GRPO turns Equation (C.1) into the marginal-preserving SDE $\mathrm{d}X_t = b_\theta\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t$. If p_t is the ODE marginal density, Fokker–Planck matching gives

$$b_\theta(x,t) = v_\theta(x,t) + \frac{\sigma_t^2}{2}\nabla \log p_t(x). \tag{C.2}$$

For the linear interpolant Equation (3.1), the score identity is

$$\nabla \log p_t(x) = -\frac{x + (1-t)v_t(x)}{t}, \tag{C.3}$$

so substituting the estimated velocity yields the working drift

$$b_\theta(x,t) = v_\theta(x,t) - \frac{\sigma_t^2}{2t}\big(x + (1-t)v_\theta(x,t)\big). \tag{C.4}$$

The minus sign comes from the negative score in Equation (C.3). Flow-GRPO commonly uses the schedule

$$\sigma_t = a\sqrt{\tfrac{t}{1-t}}, \qquad \sigma_t^2 = a^2\tfrac{t}{1-t}, \tag{C.5}$$

which is small near the data endpoint $t = 0$ and large near the noise endpoint $t = 1$.

C.1 Endpoint scaling of the path variance

Let $\Delta(t_k) = \mathbb{E}[\|b_\theta(X_k, t_k) - b_{\text{old}}(X_k, t_k)\|^2]$. For scalar σ_k, the mean path variance is

$$\bar{\lambda}_k = \frac{\Delta(t_k)}{\sigma_k^2}\Delta t_k, \tag{C.6}$$

and with the schedule Equation (C.5),

$$\bar{\lambda}_k = \frac{\Delta(t_k)(1-t_k)}{a^2 t_k}\Delta t_k. \tag{C.7}$$

The low-noise endpoint is therefore problematic when $\Delta(t)/t$ does not vanish: if $\Delta(t)$ is bounded below near $t = 0$ then $\bar{\lambda}_k$ diverges as $t_k \to 0$, whereas $\Delta(t) = O(t)$ keeps the endpoint contribution finite. The governing diagnostic is the variance-normalized drift scale $\Delta(t)/\sigma_t^2$.

D Derivations and proofs

This appendix collects the derivations and proofs for the finite-grid path-ratio law of Section 3.2.

Proof of Proposition 3.2. The kernels are Gaussian with the same covariance $\Sigma_k \Delta t_k = \sigma_k \sigma_k^\top \Delta t_k$ and means $x_k + b_\theta \Delta t_k$ and $x_k + b_{\text{old}} \Delta t_k$. The determinant terms in the Gaussian densities cancel. Expanding the difference of the two quadratic forms and using $X_{k+1} - X_k - b_{\text{old}}\Delta t_k = \sigma_k \Delta W_k$ gives

$$\xi_k = \left\langle \sigma_k^{-1}\delta b_k, \Delta W_k \right\rangle - \frac{1}{2}\left\|\sigma_k^{-1}\delta b_k\right\|^2 \Delta t_k, \tag{D.1}$$

which is Equation (3.4). □

D.1 Why the path-variance scalar takes this form

Start with the scalar case. Let the old transition be $p_{\text{old}} = \mathcal{N}(\mu_{\text{old}}, s^2)$ and the new transition be $p_\theta = \mathcal{N}(\mu_\theta, s^2)$ with the same variance. Write the mean shift in units of the standard deviation as

$$a = \frac{\mu_\theta - \mu_{\text{old}}}{s}. \tag{D.2}$$

If $X \sim p_{\text{old}}$, then $X = \mu_{\text{old}} + sZ$ with $Z \sim \mathcal{N}(0,1)$, and direct substitution into the Gaussian densities gives

$$\log \frac{p_\theta(X)}{p_{\text{old}}(X)} = aZ - \frac{1}{2}a^2. \tag{D.3}$$

Thus the same number a^2 controls both statistics of the log-ratio:

$$\mathbb{E}_{\text{old}}\left[\log \frac{p_\theta(X)}{p_{\text{old}}(X)}\right] = -\frac{1}{2}a^2, \qquad \text{Var}_{\text{old}}\left(\log \frac{p_\theta(X)}{p_{\text{old}}(X)}\right) = a^2. \tag{D.4}$$

This equality is the reason for defining a path-variance scalar: it is the squared policy move measured in noise units.

The multivariate case is the same calculation after whitening the noise. Suppose $p_{\text{old}} = \mathcal{N}(\mu_{\text{old}}, \Sigma)$ and $p_\theta = \mathcal{N}(\mu_\theta, \Sigma)$ share covariance Σ. Let

$$a = \Sigma^{-1/2}(\mu_\theta - \mu_{\text{old}}). \tag{D.5}$$

Under p_{old}, write $X = \mu_{\text{old}} + \Sigma^{1/2} Z$ with $Z \sim \mathcal{N}(0, I)$. Expanding the two quadratic forms gives

$$\log \frac{p_\theta(X)}{p_{\text{old}}(X)} = \langle a, Z\rangle - \frac{1}{2}\|a\|^2, \tag{D.6}$$

so that

$$\mathbb{E}_{\text{old}}\left[\log \frac{p_\theta(X)}{p_{\text{old}}(X)}\right] = -\frac{1}{2}\|a\|^2, \qquad \text{Var}_{\text{old}}\left(\log \frac{p_\theta(X)}{p_{\text{old}}(X)}\right) = \|a\|^2. \tag{D.7}$$

The scalar that appears from the likelihood ratio is therefore the Mahalanobis distance between the two means,

$$\|a\|^2 = (\mu_\theta - \mu_{\text{old}})^\top \Sigma^{-1} (\mu_\theta - \mu_{\text{old}}). \tag{D.8}$$

For one Euler–Maruyama Flow-GRPO step, the old and new kernels share covariance $\Sigma_k \Delta t_k = \sigma_k \sigma_k^\top \Delta t_k$ and differ only in mean by $\mu_{\theta,k} - \mu_{\text{old},k} = (b_\theta - b_{\text{old}})(X_k, t_k)\Delta t_k$. Substituting these into Equation (D.8) gives

$$(\mu_{\theta,k} - \mu_{\text{old},k})^\top (\Sigma_k \Delta t_k)^{-1} (\mu_{\theta,k} - \mu_{\text{old},k}) = \left\|\sigma_k^{-1}\big(b_\theta - b_{\text{old}}\big)(X_k, t_k)\right\|^2 \Delta t_k = \lambda_k, \tag{D.9}$$

so the one-step log-ratio has mean $-\lambda_k/2$ and variance λ_k, and λ_k is the scalar the Gaussian likelihood ratio itself uses for its drift and variance.

Proof of Proposition 3.3. Conditioned on $\mathcal{F}_{t_k}$, h_k is fixed and $\Delta W_k \sim \mathcal{N}(0, \Delta t_k I)$. Hence $\langle h_k, \Delta W_k \rangle \sim \mathcal{N}(0, \|h_k\|^2 \Delta t_k) = \mathcal{N}(0, \lambda_k)$. Subtracting $\lambda_k/2$ gives Equation (3.6). Since $r_k = \exp(\xi_k)$, r_k is lognormal. If $Z \sim \mathcal{N}(\mu, s^2)$, then $\mathbb{E}[e^Z] = e^{\mu + s^2/2}$, $\text{median}(e^Z) = e^\mu$, and $\text{Var}(e^Z) = e^{2\mu+s^2}(e^{s^2} - 1)$. Plugging $\mu = -\lambda_k/2$ and $s^2 = \lambda_k$ gives Equation (3.7). □

Proof of Proposition 3.4. Taking the conditional expectation of Equation (3.8) removes the noise term and leaves $-(2D)^{-1} \sum_m a_{k,m}^2 = -\lambda_k^{\text{center}}/2$. Since the $\epsilon_{k,m}$ are independent with unit variance,

$$\text{Var}(\bar{\xi}_k \mid \mathcal{F}_{t_k}) = \frac{1}{D^2} \sum_{m=1}^{D} a_{k,m}^2 = \lambda_k^{\text{var}}. \tag{D.10}$$

□

Corollary D.1 (Total variance for reduced log-ratios)**.** *For mean-reduced latent log-probabilities, the batch variance combines conditional transition noise with heterogeneity in the conditional centers,* $\text{Var}(\bar{\xi}_k) = \mathbb{E}[\lambda_k^{\text{var}}] + \frac{1}{4}\text{Var}(\lambda_k^{\text{center}})$.

Proof of Corollary D.1. Apply the law of total variance to Proposition 3.4:

$$\text{Var}(\bar{\xi}_k) = \mathbb{E}[\text{Var}(\bar{\xi}_k \mid \mathcal{F}_{t_k})] + \text{Var}(\mathbb{E}[\bar{\xi}_k \mid \mathcal{F}_{t_k}]) \tag{D.11}$$

$$= \mathbb{E}[\lambda_k^{\text{var}}] + \text{Var}(-\lambda_k^{\text{center}}/2). \tag{D.12}$$

□

The same argument gives the unconditional moments of the full log-ratio, which are useful when comparing against batch statistics that mix many conditional states.

Corollary D.2 (Unconditional first and second moments of log-ratios)**.** *Let* $\bar{\lambda}_k = \mathbb{E}[\lambda_k]$. *Then*

$$\mathbb{E}[\xi_k] = -\frac{1}{2}\bar{\lambda}_k, \qquad \text{Var}(\xi_k) = \bar{\lambda}_k + \frac{1}{4}\text{Var}(\lambda_k) \geq \bar{\lambda}_k. \tag{D.13}$$

Proof. The expectation follows by the tower property. For the variance, use the law of total variance,

$$\text{Var}(\xi_k) = \mathbb{E}[\text{Var}(\xi_k \mid \mathcal{F}_{t_k})] + \text{Var}(\mathbb{E}[\xi_k \mid \mathcal{F}_{t_k}]) = \mathbb{E}[\lambda_k] + \text{Var}(-\lambda_k/2). \tag{D.14}$$

□

E Analytic ratio calibration

This section gives the longer derivation of LambdaNorm-T, the analytic ratio calibration used by the main method in Section 3.3. The question is whether the reduced-log-ratio law can replace empirical RatioNorm itself. We call the construction *LambdaNorm-T*, where "T" denotes a target log-ratio variance.

Centering alone is not enough. If

$$y_{i,k} := \overline{\log r}_{i,k}, \tag{E.1}$$

then the reduced law gives

$$\mathbb{E}[y_{i,k} \mid \mathcal{F}_{t_k}] = -\frac{1}{2}\lambda_{i,k}^{\text{center}}, \qquad \text{Var}(y_{i,k} \mid \mathcal{F}_{t_k}) = \lambda_{i,k}^{\text{var}}. \tag{E.2}$$

The reduced ratio $\exp(y_{i,k})$ is therefore a surrogate ratio rather than the full path Radon–Nikodym derivative, and it generally does not preserve the mean-one likelihood-ratio convention:

$$\mathbb{E}[\exp(y_{i,k}) \mid \mathcal{F}_{t_k}] = \exp\left(-\frac{1}{2}\lambda_{i,k}^{\text{center}} + \frac{1}{2}\lambda_{i,k}^{\text{var}}\right). \tag{E.3}$$

The analytically standardized coordinate is therefore

$$z_{i,k} = \frac{y_{i,k} + \frac{1}{2}\lambda_{i,k}^{\text{center}}}{\sqrt{\lambda_{i,k}^{\text{var}} + \varepsilon}}. \tag{E.4}$$

Under the Gaussian transition approximation, $z_{i,k}$ is approximately standard normal. Using $\exp(z_{i,k})$ directly would impose an arbitrary and often too-large surrogate ratio scale. LambdaNorm-T instead chooses a target log-ratio variance $\nu_k > 0$ and reconstructs a calibrated surrogate log-ratio:

$$\widetilde{\ell}_{i,k} = \sqrt{\nu_k}\, z_{i,k} - \frac{1}{2}\nu_k, \qquad \widetilde{r}_{i,k} = \exp(\widetilde{\ell}_{i,k}). \tag{E.5}$$

If $z_{i,k} \sim \mathcal{N}(0,1)$, then $\widetilde{\ell}_{i,k} \sim \mathcal{N}(-\nu_k/2, \nu_k)$ and $\mathbb{E}[\widetilde{r}_{i,k}] \approx 1$. The subtraction $-\nu_k/2$ is important: it preserves the mean-one ratio convention after choosing the PPO-scale variance.

The target variance ν_k is tied to the PPO clipping scale rather than tuned as an unconstrained normalization constant. If the clipped objective uses a ratio window $[1-\epsilon, 1+\epsilon]$, the corresponding positive log-ratio boundary is $\log(1+\epsilon)$. We therefore use

$$\nu_k = \log(1+\epsilon)^2 \approx \epsilon^2 \tag{E.6}$$

as the default scale, with small sensitivity sweeps around it. For the standard PPO-scale choice $\epsilon = 0.2$, this gives $\nu \approx 0.04$, i.e. a target surrogate log-ratio standard deviation of about 0.2.

Equivalently, define the centered reduced log-ratio

$$\ell_{i,k}^{c} = y_{i,k} + \frac{1}{2}\lambda_{i,k}^{\text{center}}. \tag{E.7}$$

Then LambdaNorm-T can be written as

$$\widetilde{\ell}_{i,k} = T_{i,k}\ell_{i,k}^{c} - \frac{1}{2}T_{i,k}^2\lambda_{i,k}^{\text{var}}, \qquad T_{i,k} = \sqrt{\frac{\nu_k}{\lambda_{i,k}^{\text{var}} + \varepsilon}}. \tag{E.8}$$

This form makes clear that empirical RatioNorm is being replaced by analytic centering plus analytic variance calibration. In implementation we stop-gradient through λ^{center} and λ^{var} inside the normalizer, so the model cannot win by gaming the coordinate system. The budget penalty still uses differentiable λ^{center}, because that term must shape the update.

The PPO objective uses $\widetilde{r}_{i,k}$ in place of the empirical RatioNorm ratio:

$$\mathcal{L}_{\text{PPO}}^{\text{LambdaNormT}} = -\mathbb{E}_{i,k}\left[\min\left(\widetilde{r}_{i,k}A_i, \text{clip}(\widetilde{r}_{i,k}, 1-\epsilon, 1+\epsilon)A_i\right)\right]. \tag{E.9}$$

The main method combines this calibrated surrogate ratio with the damp-only λ-gradient weight from Equation (3.12). The earlier primal-dual ablation instead adds the exponential-moving-average (EMA) smoothed budget penalty from Section G.1:

$$\mathcal{L} = \mathcal{L}_{\text{PPO}}^{\text{LambdaNormT}} + \beta_{\lambda,t}\phi_\alpha(\Lambda_{\text{late}} - \tau), \qquad \Lambda_{\text{late}} = \frac{1}{|\mathcal{K}_{\text{late}}|}\sum_{k\in\mathcal{K}_{\text{late}}} \mathbb{E}_i[\lambda_{i,k}^{\text{center}}]. \tag{E.10}$$

There is one further scale issue. For a Gaussian transition $X_{k+1} \mid X_k \sim \mathcal{N}(\mu_{\theta,k}, s_k^2 I)$, the score with respect to the mean scales like

$$\nabla_\mu \log p(X_{k+1} \mid X_k) = \frac{X_{k+1} - \mu_{\theta,k}}{s_k^2}, \qquad \|\nabla_\mu \log p\| \text{ scales like } s_k^{-1}. \tag{E.11}$$

Thus low-noise timesteps can receive disproportionately large gradients. A simple analytic compensation is an optional per-step loss weight

$$w_k^{\text{score}} = \text{clip}\left(\frac{s_k}{s_{\text{ref}}}, w_{\min}, w_{\max}\right), \qquad s_k = \mathbb{E}[\sigma_k \sqrt{\Delta t_k}], \tag{E.12}$$

which is applied to the per-step PPO loss. This is an ablation, not part of the minimal LambdaNorm-T definition.

The initial LambdaNorm-T ablation grid varies around the clipping-derived central value:

$$\nu \in \{0.01,\ 0.025,\ 0.04,\ 0.09\}, \qquad \sqrt{\nu} \in \{0.10,\ 0.158,\ 0.20,\ 0.30\}. \tag{E.13}$$

The central setting is $\nu = 0.04$, with tighter and looser alternatives included to test sensitivity. We also include a score-scale-weighted variant at the same target variance.

LambdaNorm-T answers the sharper analytic question: can the finite-grid likelihood law replace most or all of empirical RatioNorm? The primary comparison reports reward together with path-variance spend:

$$R_{\text{LambdaNormT+dual}} \gtrsim R_{\text{RatioNorm}} \qquad \text{and} \qquad \Lambda_{\text{late,LambdaNormT+dual}} < \Lambda_{\text{late,RatioNorm}}. \tag{E.14}$$

This comparison tests whether analytic centering and variance calibration recover the stabilizing effect of empirical RatioNorm while preserving the same λ-budgeted control variable.

F Diagnostic and alternative estimators of the path-variance scalar

The main method uses the transition mean-shift estimator in Section 3.3. The estimators below are useful for diagnostics, sanity checks, and implementation variants, but they are not used to train the main SD3.5 hard-OCR result.

F.1 Velocity-only simplification of the drift

Often the model naturally exposes v_θ, not b_θ. If both policies use the same scalar schedule σ_t and the marginal-preserving drift Equation (C.4), then

$$b_\theta(x,t) - b_{\text{old}}(x,t) = \left[1 - \frac{\sigma_t^2(1-t)}{2t}\right](v_\theta(x,t) - v_{\text{old}}(x,t)). \tag{F.1}$$

Therefore

$$\widehat{\lambda}_{i,k}^{\text{vel}} = \left[1 - \frac{\sigma_k^2(1-t_k)}{2t_k}\right]^2 \frac{\left\|v_\theta(x_k^i, t_k, c_i) - v_{\text{old}}(x_k^i, t_k, c_i)\right\|^2}{\sigma_k^2} \Delta t_k. \tag{F.2}$$

For the schedule $\sigma_t^2 = a^2 t/(1-t)$, the bracket simplifies to $1 - a^2/2$, so

$$\widehat{\lambda}_{i,k}^{\text{vel}} = \left(1 - \frac{a^2}{2}\right)^2 \frac{\left\|v_\theta(x_k^i, t_k, c_i) - v_{\text{old}}(x_k^i, t_k, c_i)\right\|^2}{a^2 t_k/(1-t_k)} \Delta t_k. \tag{F.3}$$

This identity is useful only when applied with the same time convention and same SDE drift used by the sampler.

F.2 Ratio-based validation

The theorem also predicts

$$\mathbb{E}[\log r_k] \approx -\widehat{\lambda}_k/2, \qquad \text{Var}(\log r_k) \approx \widehat{\lambda}_k. \tag{F.4}$$

Thus one can estimate λ_k from empirical log-ratios:

$$\widehat{\lambda}_k^{\text{var}} = \widehat{\text{Var}}_i(\log r_{i,k}), \qquad \widehat{\lambda}_k^{\text{mean}} = -2\widehat{\mathbb{E}}_i[\log r_{i,k}]. \tag{F.5}$$

For mean-reduced latent log-probabilities, use the two-component validation from Proposition 3.4 instead:

$$-2\widehat{\mathbb{E}}_i[\overline{\log r}_{i,k}] \approx \widehat{\lambda}_k^{\text{center}}, \qquad \widehat{\text{Var}}_i(\overline{\log r}_{i,k}) \approx \widehat{\lambda}_k^{\text{var}}. \tag{F.6}$$

These ratio-based estimators are the convention used in the law-audit plots in Section 4. They are noisier than the transition mean-shift estimator, but they are a stringent test of the theory: if the finite-grid law is explanatory, then the mean-shift estimate and the observed log-ratio statistics should align up to sampling noise.

G Primal-dual baseline and diagnostic clipping

This appendix formalizes the two constructions used outside the main method: the primal-dual ablation reported as a baseline in Tables 2 and 3, and the diagnostic z-space normalization used in the Tiny-SD3 law-audit intervention in Section 4. Both are referenced from the main text; neither is part of the λ-Controlled GRPO algorithm.

G.1 Exponentially smoothed primal-dual budget

The primal-dual ablation constrains the late-step path-variance mean

$$\Lambda_{\text{late}} = \frac{1}{|\mathcal{K}_{\text{late}}|} \sum_{k \in \mathcal{K}_{\text{late}}} \widehat{\lambda}_k^{\text{center}}, \tag{G.1}$$

through

$$\max_{\theta} \; J(\theta) \qquad \text{subject to} \qquad \Lambda_{\text{late}} \le \tau. \tag{G.2}$$

With empirical RatioNorm as $J(\theta)$, the ablation uses the smooth hinge

$$\phi_\alpha(u) = \left[\frac{\text{softplus}(\alpha u)}{\alpha}\right]^2 \tag{G.3}$$

and primal loss

$$\mathcal{L}_t(\theta) = \mathcal{L}_{\text{GRPO/RatioNorm}} + \beta_{\lambda,t}\phi_\alpha(\Lambda_{\text{late}}(\theta) - \tau)\,. \tag{G.4}$$

The multiplier is updated outside autograd from a detached EMA:

$$\bar{\lambda}_{\text{late},t} = (1-\gamma)\bar{\lambda}_{\text{late},t-1} + \gamma\,\Lambda_{\text{late}}(\theta_t)^{\text{sg}}, \tag{G.5}$$

$$\beta_{\lambda,t+1} = \Pi_{[0,\beta_{\max}]}\left[\beta_{\lambda,t} + \eta_\beta\left(\bar{\lambda}_{\text{late},t} - \tau - \delta\right)\right]. \tag{G.6}$$

This ablation established that λ can actively control optimization, but the main method replaces the multiplier dynamics with direct damp-only λ-gradient weighting.

G.2 Diagnostic normalization and clipping

The reduced law gives the analytic diagnostic

$$z_{i,k} = \frac{\overline{\log r}_{i,k} + \widehat{\lambda}_{i,k}^{\text{center}}/2}{\sqrt{\widehat{\lambda}_{i,k}^{\text{var}} + \varepsilon}}. \tag{G.7}$$

One can either log $z_{i,k}$ as a law audit or clip it,

$$\tilde{z}_{i,k} = \text{clip}(z_{i,k}, -c, c), \tag{G.8}$$

and reconstruct

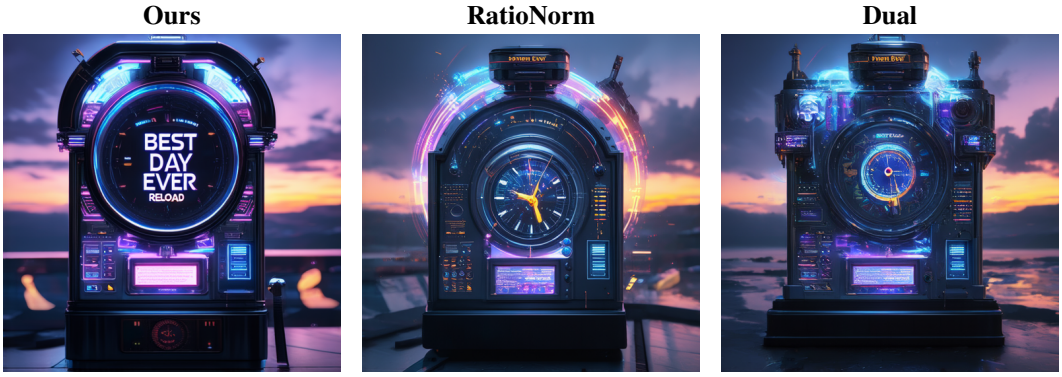
$$\widetilde{\overline{\log r}}_{i,k} = -\frac{1}{2}\widehat{\lambda}_{i,k}^{\text{center}} + \tilde{z}_{i,k}\sqrt{\widehat{\lambda}_{i,k}^{\text{var}} + \varepsilon}. \tag{G.9}$$

This implies the adaptive interval

$$\overline{\log r}_{i,k} \in \left[-\frac{1}{2}\widehat{\lambda}_{i,k}^{\text{center}} - c\sqrt{\widehat{\lambda}_{i,k}^{\text{var}}}, -\frac{1}{2}\widehat{\lambda}_{i,k}^{\text{center}} + c\sqrt{\widehat{\lambda}_{i,k}^{\text{var}}}\right]. \tag{G.10}$$

H QUALITATIVE HELD-OUT OCR COMPARISONS

The aggregate held-out metrics in Table 2 are supported by a qualitative sweep on the same 762-prompt test complement. The examples below use matched prompt indices and matched random seeds across methods: empirical RatioNorm at step 40, the primal-dual ablation at step 80, and LambdaNorm-T with damp-only λ weights at step 80. All examples are drawn from the disjoint 762-prompt held-out complement and were not used for checkpoint selection. The first block shows clear transcript-level wins for λ-Controlled GRPO. The second block deliberately includes mixed and failure cases, because the method improves the distribution of outcomes but does not solve text rendering uniformly.



Prompt. A futuristic time machine with a sleek, metallic finish, its dial set to "Best Day Ever Reload". The machine is surrounded by a glowing aura, with digital readouts and holographic interfaces displaying vibrant colors, set against a backdrop of a twilight sky.

Figure 3: Qualitative held-out test comparison for prompt 319. Target text: `Best Day Ever Reload`. λ-Controlled GRPO renders the complete target in the central display, while RatioNorm and the primal-dual ablation produce the object and lighting but fail to place readable target text.



Prompt. A bustling amusement park with a vibrant entrance arch prominently displaying "Height Limit 48in", surrounded by excited children and their parents, with colorful banners and playful music in the background.

Figure 4: Qualitative held-out test comparison for prompt 107. Target text: `Height Limit 48in`. The analytic method renders the full target on the amusement-park arch. RatioNorm and the primal-dual ablation retain the colorful park scene, but their sign text corrupts the numerical suffix or middle word.



Prompt. A charming bakery window with a vintage wooden frame, adorned with a decal that reads "Fresh Daily" in elegant cursive. Sunlight streams through, casting a warm glow on the display of freshly baked bread and pastries.

Figure 5: Qualitative held-out test comparison for prompt 145. Target text: `Fresh Daily`. The analytic method gives the cleanest bakery window composition and a readable decal, with only a small character-level artifact in "Fresh." RatioNorm repeats the text, and the primal-dual output is more typographically fragmented.



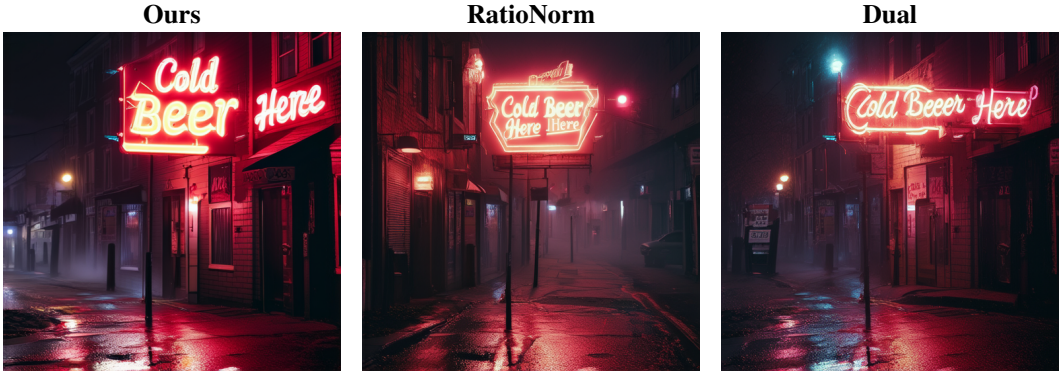
Prompt. A bustling school cafeteria on a Friday, with a large, colorful sign displaying the "Pizza Friday Special" menu. Students in vibrant uniforms gather excitedly, pointing at the mouth-watering pizzas arranged on the serving counter.

Figure 6: Qualitative held-out test comparison for prompt 366. Target text: `Pizza Friday Special`. λ-Controlled GRPO writes the full cafeteria banner cleanly. RatioNorm and the primal-dual ablation preserve the school-lunch scene, but their banner text is partial or corrupted.



Prompt. An astronaut sits at a desk inside Moon Base Alpha, writing in a journal. The base's futuristic interior is illuminated by soft blue lights, and a large window behind the astronaut showcases the barren lunar landscape and Earth rising above the horizon. "Moon Base Alpha" is prominently displayed on a plaque nearby.

Figure 7: Qualitative held-out test comparison for prompt 747. Target text: `Moon Base Alpha`. λ-Controlled GRPO keeps the full phrase readable on the base sign. The baselines generate coherent lunar-base scenes, but their signage repeats or corrupts the target phrase.



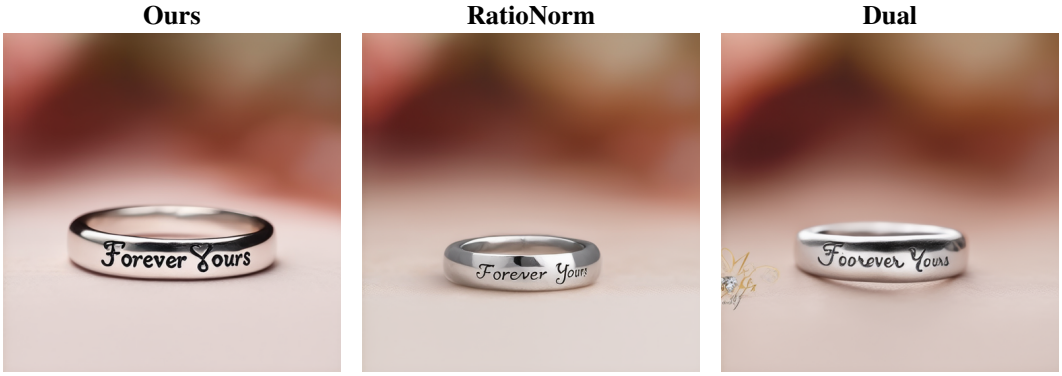
Prompt. A gritty urban street at night, with a dive bar neon sign glowing brightly, reading "Cold Beer Here", casting a warm, inviting glow through the misty air, reflecting off the wet pavement.

Figure 8: Qualitative held-out test comparison for prompt 168. Target text: `Cold Beer Here`. The analytic method produces the most legible target phrase and strongest neon-street atmosphere. The remaining error is spatial: "Here" is readable but spills onto the adjacent facade rather than remaining entirely within the main sign.



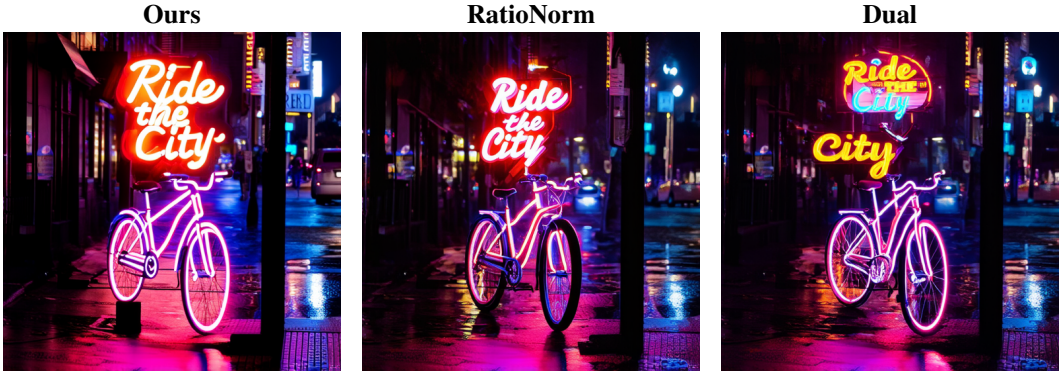
Prompt. A serene desert scene with a weathered signpost pointing towards "Hidden Water Spring", surrounded by tall palm trees and golden sand dunes, under a clear blue sky.

Figure 9: Qualitative held-out test comparison for prompt 176. Target text: `Hidden Water Spring`. The analytic method gives an exact, centered, readable sign in the intended desert scene. The comparison highlights the kind of sample counted by the stricter exact and substring metrics rather than only by mean OCR reward.



Prompt. A close-up photograph of an engraved silver ring with the inscription "Forever Yours" delicately etched into its surface, set against a soft, blurred background of romantic, warm tones.

Figure 10: Qualitative held-out test comparison for prompt 002. Target text: `Forever Yours`. Both λ-Controlled GRPO and RatioNorm render this short phrase cleanly; the primal-dual ablation introduces a character-level error.



Prompt. A neon bike rental sign glowing "Ride the City" stands out against a dark urban backdrop, its vibrant colors reflecting off wet pavements in a bustling night scene.

Figure 11: Qualitative held-out test comparison for prompt 047. Target text: `Ride the City`. λ-Controlled GRPO preserves the complete phrase, while the baselines tend to omit or repeat words.



Prompt. A glowing Magic 8 Ball floats in a dimly lit room, its triangular window displaying the answer "Ask Again Later" in shimmering, ethereal blue text, surrounded by a soft, mystical aura.

Figure 12: Qualitative held-out test comparison for prompt 130. Target text: `Ask Again Later`. The proposed method keeps the three-word answer legible inside the triangular display; the baselines introduce spelling noise.



Prompt. A realistic photograph of a vintage vending machine with a prominent sign that reads "Exact Change Only", set against a slightly worn brick wall, with a few coins scattered at its base.

Figure 13: Qualitative held-out test comparison for prompt 321. Target text: `Exact Change Only`. λ-Controlled GRPO places the phrase on the vending machine sign, while the baselines produce less recognizable signage.



Prompt. A futuristic space hotel lobby with a sleek, glowing sign that reads "ZeroG Suites", surrounded by minimalist decor and large windows showcasing the vastness of space outside.

Figure 14: Qualitative held-out test comparison for prompt 456. Target text: `ZeroG Suites`. The proposed method renders the hotel sign cleanly; the baselines are close but introduce extra or incorrect letters.



Prompt. A vast, icy landscape with a stark, metal sign reading "Titanic Memorial Site" standing firmly on a snow-covered rock near a towering iceberg, the cold waters of the North Atlantic stretching endlessly into the horizon.

Figure 15: Qualitative held-out test comparison for prompt 654. Target text: `Titanic Memorial Site`. λ-Controlled GRPO preserves the full memorial sign text in the icy scene, while the baselines retain only fragments.



Prompt. A realistic photograph of a handwritten note, with the words "Eat Healthy Today" clearly visible, taped to the door of a modern refrigerator in a well-lit kitchen.

Figure 16: Qualitative held-out test comparison for prompt 706. Target text: `Eat Healthy Today`. The proposed method keeps the note text readable; the baselines drop or corrupt parts of the phrase.

H.1 MIXED AND FAILURE CASES

The examples above show where the analytic intervention is most visually clear. For calibration, we also scan the disjoint held-out test complement for prompts where the proposed method is comparable to RatioNorm, worse than RatioNorm, or where all methods remain far from the requested text. These cases are useful scientifically: they separate an average improvement in transcript success from a claim of uniform dominance.



Prompt. A vast, empty city square with a large, blank billboard standing prominently, awaiting the words "Your Ad Here" to be filled in, under a clear blue sky with a few fluffy clouds.

Figure 17: Mixed held-out test comparison for prompt 718. Target text: `Your Ad Here`. This is a clear negative case for the proposed method: λ-Controlled GRPO leaves the billboard effectively blank, while both empirical RatioNorm and the primal-dual ablation place the intended phrase cleanly in the center.



Prompt. A detailed amusement park map with colorful pathways and attractions, prominently marking "You Are Here" with a large, red arrow. The map is held by a friendly park employee, standing in front of a vibrant, bustling entrance.

Figure 18: Mixed held-out test comparison for prompt 617. Target text: `You Are Here`. The proposed method renders the main phrase but adds extra corrupted text, which hurts the transcript metric. RatioNorm and the primal-dual ablation give cleaner target text in this particular sample.



Prompt. A high-tech laboratory setting with a test tube labeled "Sample XZ42" on a sleek, illuminated stand, surrounded by advanced scientific equipment and glowing monitors displaying complex data. The scene is modern and sterile, with a scientist in the background observing through a protective visor.

Figure 19: Failure held-out test comparison for prompt 026. Target text: `Sample XZ42`. All three methods produce plausible laboratory imagery, but none renders the alphanumeric label correctly. This illustrates a residual failure mode for short technical strings and small curved surfaces.



Prompt. Retro diner scene with a red and white checkered placemat featuring the text "Todays Special Atomic Burger" in bold, vintage font. The placemat is slightly worn, with a classic 1950s diner background.

Figure 20: Mixed held-out test comparison for prompt 159. Target text: `Todays Special Atomic Burger`. This is a case where empirical RatioNorm is better: it keeps the main phrase nearly intact, while λ-Controlled GRPO produces a plausible diner placemat but scrambles the target into extra and reordered words.



Prompt. "Skate or Die" slogan prominently displayed on a vibrant, colorful skateboard deck, set against a backdrop of a bustling urban skate park at sunset, with skaters in motion and graffiti-covered walls, capturing the rebellious spirit and dynamic energy of skate culture.

Figure 21: Mixed held-out test comparison for prompt 054. Target text: `Skate or Die`. All three samples are visually plausible skateboard-deck renderings, but this is not a clean win for the proposed method: the OCR system fails on the stylized version, while the RatioNorm and primal-dual samples are closer to the normalized transcript.



Prompt. A weathered treasure map laid out on an old wooden table, with "X Marks the Spot" clearly visible in the center, surrounded by intricate illustrations of mountains, forests, and a distant coastline, all under the warm glow of a vintage lamp.

Figure 22: Failure held-out test comparison for prompt 040. Target text: `X Marks the Spot`. All methods capture the map-and-lamp composition, but none renders the complete target phrase. This suggests that some failures are driven by scene/text placement difficulty rather than only by ratio instability.

Table 6 reports the per-checkpoint PickScore validation trajectory that underlies the winner selection in Table 4.

Method	Ckpt.	Mean PickScore ↑	Std.	Λ_{late} ↓
RatioNorm	40	0.8325	0.057	3.8×10^{-5}
RatioNorm	60	0.8266	0.057	6.4×10^{-5}
RatioNorm	80	0.8126	0.059	6.80×10^{-3}
λ-Controlled GRPO	40	0.8365	0.057	4.65×10^{-3}
λ-Controlled GRPO	60	0.8373	0.055	1.9×10^{-4}
λ-Controlled GRPO	80	**0.8380**	0.057	3.34×10^{-3}

Table 6: Random 256-prompt SD3.5 PickScore validation comparison. Every λ-Controlled GRPO checkpoint beats every RatioNorm checkpoint. RatioNorm peaks at step 40 and then regresses as its late-step λ spend jumps past $\tau \approx 2.0\times10^{-3}$ by step 80, an overspend of about 3.4τ. The analytic method improves PickScore monotonically across $40\to60\to80$ and keeps Λ_{late} at most comparable to τ.

I Qualitative held-out PickScore comparisons

The PickScore aggregate result in Table 4 is illustrated by a qualitative sweep on the same 762-prompt held-out test complement. Comparisons are matched by prompt index and random seed across methods: empirical RatioNorm at step 40 (the validation-favored early stop) and λ-Controlled GRPO at step 80 (the validation-favored late stop). Every example below is from the disjoint 762-prompt held-out complement and was not used for checkpoint selection. Captions report the prompt and the raw PickScore values verbatim, so the reader can judge each pair. The first block shows PickScore wins for the analytic method, the second block shows near-ties or PickScore losses, and the third block shows cases where both methods score low on the reward model.

I.1 PickScore hits (analytic method wins)

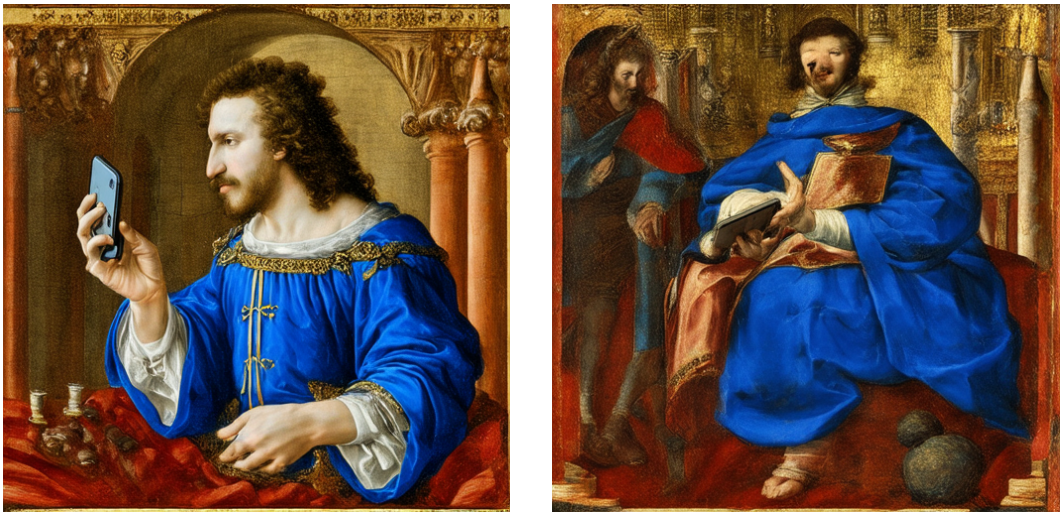
λ-Controlled GRPO **RatioNorm**



Prompt. anime girl in the space with a sign saying ‘sdxl’ as text,4k

Figure 23: Held-out test prompt 274. PickScore: Ours $=$ 0.9137, RatioNorm $=$ 0.8235, Δ $=$ $+0.0902$.

λ-Controlled GRPO **RatioNorm**



Prompt. medieval painting of a man in a blue gown using a cellphone

Figure 24: Held-out test prompt 521. PickScore: Ours $= 0.9014$, RatioNorm $= 0.8180$, $\Delta = +0.0834$.

λ-Controlled GRPO **RatioNorm**



Prompt. portrait of a giraffe in a fiery thunderstorm, digital art, hyper detailed

Figure 25: Held-out test prompt 853. PickScore: Ours $= 0.9170$, RatioNorm $= 0.8512$, $\Delta = +0.0658$.

λ-Controlled GRPO **RatioNorm**



Prompt. Working in a diamond mine, Midjourney v5 style, insanely detailed, photorealistic, 8k, volumetric lighting

Figure 26: Held-out test prompt 223. PickScore: Ours $= 0.8623$, RatioNorm $= 0.7737$, $\Delta = +0.0886$.

λ-Controlled GRPO **RatioNorm**



Prompt. Cyberpunk, Ancient India style, sci fi, silver on a black background, bas-relief, cyborgs, neon lighting, contrasting shadows, three-dimensional sculpture, high resolution, 8k detail, baroque, clear edges, technology, mechanisms

Figure 27: Held-out test prompt 682. PickScore: Ours $= 0.8182$, RatioNorm $= 0.7329$, $\Delta = +0.0853$.

λ-Controlled GRPO **RatioNorm**



Prompt. Synesthesia, musical notes flying away from a music flower, charcoal hyperrealistic colourful art print by robert longo and alicexz

Figure 28: Held-out test prompt 626. PickScore: Ours $= 0.8725$, RatioNorm $= 0.7951$, $\Delta = +0.0774$.

λ-Controlled GRPO **RatioNorm**



Prompt. A set of museum-quality emerald bracelets and beads in green in a display box at the auction, 32k, highest resolution, hyper realistic

Figure 29: Held-out test prompt 27. PickScore: Ours $= 0.8773$, RatioNorm $= 0.8050$, $\Delta = +0.0723$.

λ-Controlled GRPO **RatioNorm**



Prompt. film still of Neytiri from Avatar

Figure 30: Held-out test prompt 118. PickScore: Ours $= 0.8441$, RatioNorm $= 0.7749$, $\Delta = +0.0692$.

λ-Controlled GRPO **RatioNorm**



Prompt. rutger hauer from blade runner standing in the rain, green light, vhs quality, film grain, wet, very sad and reluctant expression, wearing a biomechanical suit, scifi, digital painting, artstation, concept art

Figure 31: Held-out test prompt 835. PickScore: Ours $= 0.8748$, RatioNorm $= 0.8071$, $\Delta = +0.0677$.

λ-Controlled GRPO **RatioNorm**



Prompt. a portal to snowy mountain, standing in a warm summer field

Figure 32: Held-out test prompt 1003. PickScore: Ours $= 0.9609$, RatioNorm $= 0.8936$, $\Delta = +0.0673$.

λ-Controlled GRPO **RatioNorm**



Prompt. art print of a cute fire elemental pokemon by league of legends. finally evolutionary stage

Figure 33: Held-out test prompt 514. PickScore: Ours $= 0.9264$, RatioNorm $= 0.8594$, $\Delta = +0.0670$.

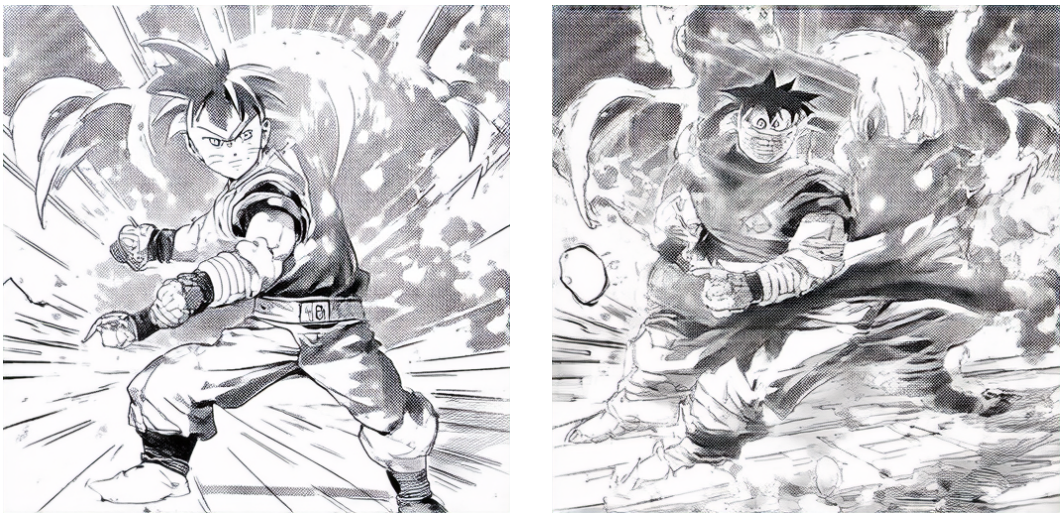
λ-Controlled GRPO **RatioNorm**



Prompt. Nissan GT-R in a parking lot, raining, film grain, moody

Figure 34: Held-out test prompt 149. PickScore: Ours $= 0.8943$, RatioNorm $= 0.8288$, $\Delta = +0.0655$.

λ-Controlled GRPO **RatioNorm**



Prompt. a manga drawing of naruto fighting goku

Figure 35: Held-out test prompt 130. PickScore: Ours $= 0.8538$, RatioNorm $= 0.7916$, $\Delta = +0.0622$.

λ-Controlled GRPO **RatioNorm**



Prompt. Movie still of star wars young luke skywalker working as mechanic in a garage, extremely detailed, intricate, high resolution, hdr, trending on artstation

Figure 36: Held-out test prompt 852. PickScore: Ours $= 0.8465$, RatioNorm $= 0.7915$, $\Delta = +0.0550$.

λ-Controlled GRPO **RatioNorm**



Prompt. photo of king kong lifting a landrover defender in the jungle river, misty mud rocks, headlights chrome detailing

Figure 37: Held-out test prompt 307. PickScore: Ours $= 0.8618$, RatioNorm $= 0.8076$, $\Delta = +0.0542$.

I.2 PICKSCORE NEAR-TIES AND RATIONORM WINS

λ-Controlled GRPO **RatioNorm**



Prompt. analog style picture of a lizard dressed as a knight in armour

Figure 38: Held-out test prompt 942. PickScore: Ours $= 0.8598$, RatioNorm $= 0.9018$, $\Delta = -0.0420$.

λ-Controlled GRPO **RatioNorm**



Prompt. A fortnite map inspired by Star Wars

Figure 39: Held-out test prompt 597. PickScore: Ours $= 0.8291$, RatioNorm $= 0.8680$, $\Delta = -0.0389$.

I.3 PickScore joint failures (both methods score low)

λ-Controlled GRPO **RatioNorm**



Prompt. Young Victoria coren-mitchell, insanely detailed, photorealistic, 8k, ultra high resolution, volumetric lighting, taken with canon eos,

Figure 40: Held-out test prompt 391. PickScore: Ours $= 0.6642$, RatioNorm $= 0.6773$, $\Delta = -0.0131$. Both methods land well below the aggregate mean; this prompt class (photorealistic named-individual likeness) is a known hard case for SD3.5-LoRA regardless of the ratio controller.

λ-Controlled GRPO **RatioNorm**



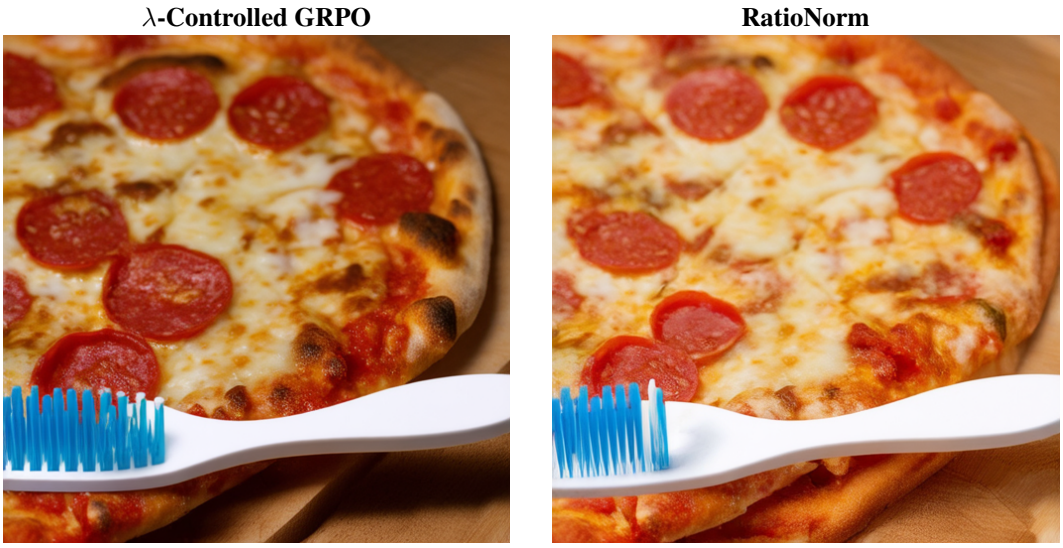
Prompt. an epic view of a demonic Rose-ringed parakeet cyborg inside an ironmaiden robot, wearing a noble robe, large view, a surrealist painting, aralan bean and Philippe Druillet, hiromu arakawa, volumetric lighting, detailed shadows

Figure 41: Held-out test prompt 10. PickScore: Ours $= 0.6661$, RatioNorm $= 0.6675$, $\Delta = -0.0015$. Heavy compositional nesting with named artist styles; both methods score near the bottom of the distribution.

J QUALITATIVE HELD-OUT GENEVAL COMPARISONS

This appendix illustrates the compositional GenEval regime with matched held-out image comparisons between the analytic λ-Controlled GRPO method and empirical RatioNorm on SD3.5. Each pair uses the same prompt and the same random seed at each method's validation-selected checkpoint, and every example is drawn from the disjoint held-out test complement that was not used for checkpoint selection. Captions report the prompt, its GenEval category tag, and the raw per-prompt GenEval and strict-accuracy scores, so the reader can judge each pair. The first block shows GenEval wins for the analytic method, the second shows joint successes where both methods pass, and the third shows cases where RatioNorm wins.

J.1 GENEVAL HITS (ANALYTIC METHOD WINS)



Prompt. a photo of a toothbrush below a pizza (`position`)

Figure 42: Held-out test prompt 117. GenEval: Ours $= 1.00$, RatioNorm $= 0.00$. Strict accuracy: Ours $= 1$, RatioNorm $= 0$.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a green couch and an orange umbrella (`color_attr`)

Figure 43: Held-out test prompt 124. GenEval: Ours $= 1.00$, RatioNorm $= 0.00$. Strict accuracy: Ours $= 1$, RatioNorm $= 0$.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a toaster below a traffic light (`position`)

Figure 44: Held-out test prompt 234. GenEval: Ours $= 1.00$, RatioNorm $= -0.33$. Strict accuracy: Ours $= 1$, RatioNorm $= 0$.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of three sinks (`counting`)

Figure 45: Held-out test prompt 423. GenEval: Ours $= 1.00$, RatioNorm $= 0.33$. Strict accuracy: Ours $= 1$, RatioNorm $= 0$.

J.2 GenEval joint successes (both methods pass)

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a cat (`single_object`)

Figure 46: Held-out test prompt 5. GenEval: Ours $= 1.00$, RatioNorm $= 1.00$. Both methods produce a strict pass; included to show that the matched-seed protocol is not exploiting failure modes of one side.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of two trains (`counting`)

Figure 47: Held-out test prompt 8. GenEval: Ours $= 1.00$, RatioNorm $= 1.00$.

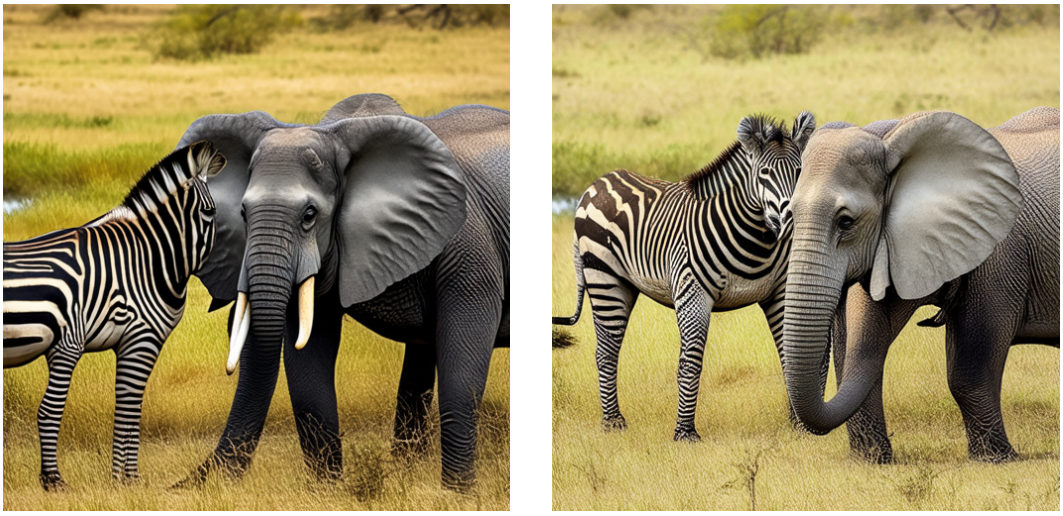
λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a yellow bird and a black motorcycle (`color_attr`)

Figure 48: Held-out test prompt 9. GenEval: Ours $= 1.00$, RatioNorm $= 1.00$.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a zebra left of an elephant (`position`)

Figure 49: Held-out test prompt 25. GenEval: Ours $= 1.00$, RatioNorm $= 1.00$. A rare `position` joint success on a category that is otherwise near-floor for both methods.

J.3 GENEVAL LOSSES (RATIONORM WINS)

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a kite above a toothbrush (`position`)

Figure 50: Held-out test prompt 39. GenEval: Ours $= 0.33$, RatioNorm $= 1.00$. Strict accuracy: Ours $= 0$, RatioNorm $= 1$.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of three baseball bats (`counting`)

Figure 51: Held-out test prompt 62. GenEval: Ours $= 0.00$, RatioNorm $= 1.00$. Strict accuracy: Ours $= 0$, RatioNorm $= 1$.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a green bus (`colors`)

Figure 52: Held-out test prompt 105. GenEval: Ours $= 0.00$, RatioNorm $= 1.00$. Strict accuracy: Ours $= 0$, RatioNorm $= 1$. The bus is recognized in both images, but only the RatioNorm rendering is judged strictly green by the GenEval color classifier.

λ-Controlled GRPO **RatioNorm**



Prompt. a photo of a pizza and a bench (`two_object`)

Figure 53: Held-out test prompt 136. GenEval: Ours $= 0.00$, RatioNorm $= 1.00$. Strict accuracy: Ours $= 0$, RatioNorm $= 1$.