# OPTIMIZING STOCHASTIC POLICY GEOMETRY FOR FEW-STEP FLOW REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Few-step flow models generate samples through a handful of large denoising transitions. Applying GRPO to such models requires introducing stochasticity into the denoising transitions, and when a single transition spans a wide range of noise levels, the choice of the resulting stochastic policy becomes a major factor in optimization. Recent constructions constrain the policy to preserve the target flow coefficients, but this constraint admits a family of policies and offers no criterion for choosing among them. We introduce Geo-FSRL (Geometry-Optimized Few-Step Reinforcement Learning), built on a selection principle: at a fixed conditioning state and timestep pair, choose the admissible policy that minimizes the conditional KL divergence induced by any given nonzero velocity-output update. Within a coefficient-preserving family of isotropic Gaussian transitions, this criterion reduces to minimizing a single scalar geometry factor that determines the conditional KL for every velocity-output update. The same factor governs velocity-space Fisher information and conditional transition log-likelihood-ratio variance. We derive the unique closed-form optimizer within this family, shared across all nonzero update magnitudes and directions, with transition coefficients determined by both the source and target noise levels. Because the minimized geometry factor still varies across timesteps, we further adopt gradient-scale matching from prior work and derive closed-form policy-gradient and KL weights that normalize explicit score scales and reference-KL curvature. Geo-FSRL trains on the original few-step schedule and retains the deterministic solver at inference. Our method achieves higher final reward scores than Flow-GRPO and Flow-CPS on both image generation and video generation models under the 4-step setting.

## 1 INTRODUCTION

Reinforcement learning (RL) is increasingly used to post-train diffusion and flow models for visual generation, improving visual quality and prompt alignment through reward feedback (Black et al., 2024; Xue et al., 2025; Liu et al., 2025). A common formulation treats denoising as a sequence of decisions, with intermediate latents as states and rewards assigned to generated samples. For deterministic flow samplers, Flow-GRPO and DanceGRPO introduce stochastic transitions to support exploration and likelihood-ratio optimization (Liu et al., 2025; Xue et al., 2025). Meanwhile, distillation compresses generation into a few denoising steps (Yin et al., 2024; Contributors, 2025), motivating RL methods that retain this efficient inference schedule. Each transition then spans a wide interval of noise levels, making the finite-step stochastic policy an important design choice.

Large transitions expose differences between stochastic constructions. The standard Euler–Maruyama transition of Flow-GRPO generally deviates from the coefficient-preserving condition considered here (Liu et al., 2025; Wang & Yu, 2025). Flow-CPS instead constructs transitions directly at the target noise level (Wang & Yu, 2025). We adopt its coefficient-preserving constraint, which fixes the sum of the squared noise coefficients while allowing different splits between predicted noise carried over from the current state and fresh Gaussian noise. Flow-CPS specifies this split through a single normalized stochasticity parameter shared across timesteps and tuned empirically. The constraint thus defines an admissible family but leaves a policy-selection question: *which transition should serve as the training policy?*



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.

We introduce Geo-FSRL (Geometry-Optimized Few-Step Reinforcement Learning), motivated by the distinction between the distributions used for policy optimization and the velocity predictions used for generation. GRPO moderates updates through likelihood-ratio clipping and reference-policy regularization, both defined on stochastic transition distributions, whereas deterministic inference depends on velocity predictions. Within the coefficient-preserving isotropic Gaussian family, the same velocity-output update can induce different changes in the transition distribution depending on the noise split. We therefore propose the following selection principle: *at a fixed conditioning state and timestep pair, choose the admissible policy that minimizes the conditional KL divergence induced by any given nonzero velocity-output update*.

To characterize this criterion, we show that the conditional KL is exactly one half of a scalar geometry factor squared times the squared velocity-update norm. The squared factor quantifies *policy sensitivity*, the normalized conditional KL response to velocity-output changes. The same factor governs velocity-space Fisher information and conditional log-likelihood-ratio variance under the old policy. This characterization holds for finite updates and yields the same minimizing coefficients for all nonzero update magnitudes and directions, before reward-based optimization determines which updates to take.

The resulting problem has a unique closed-form solution, with coefficients determined by both source and target noise levels. Under rectified-flow parameterization, it recovers the one-step SDE-DPM-Solver++ transition (Lu et al., 2022), providing a geometric optimality characterization of an existing sampling rule. At a fixed target noise level, the minimized geometry factor increases with the transition span, so larger transitions remain more sensitive to the same velocity update even under the optimal split. To account for the remaining variation across timesteps, we follow the gradient-scale matching principle of Team et al. (2025) to derive policy-gradient and KL weights that normalize explicit score scales and reference-KL curvature.

We evaluate Geo-FSRL on four-step distilled Qwen-Image (Wu et al., 2025; Contributors, 2025) and Wan2.1 (Wan et al., 2025) models for image and video generation. It attains higher final reward scores than Flow-GRPO and Flow-CPS in the main comparisons while retaining the original deterministic inference schedule.

Our contributions are summarized as follows:

- We propose selecting stochastic policies by minimizing conditional KL for a given nonzero velocity-output update, reducing the problem to scalar optimization within a coefficient-preserving isotropic Gaussian family.
- We derive the unique closed-form optimizer within this family, shared across all nonzero update magnitudes and directions, and characterize how the minimum geometry factor depends on the transition span at a fixed target noise level.
- We develop Geo-FSRL by combining the selected transition with closed-form policy-gradient and KL weights from gradient-scale matching, and demonstrate higher final reward scores than Flow-GRPO and Flow-CPS on four-step image and video models.

## 2 Preliminaries

### 2.1 Flow Matching and GRPO

We consider rectified-flow models defined by the linear probability path

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \qquad t \in [0,1]. \tag{1}$$

where $\boldsymbol{x}_0$ is a clean sample and the model $\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)$ is trained to predict the velocity $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$. Given a prediction, the clean sample and the noise are estimated as

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_t - t\boldsymbol{v}_\theta(\boldsymbol{x}_t, t), \qquad \hat{\boldsymbol{\epsilon}} = \boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t). \tag{2}$$

Sampling follows a decreasing timestep schedule $\mathcal{T} = \{t_0 = 1 > t_1 > \cdots > t_K = 0\}$ with $K$ denoising steps; for a transition between timesteps $t > s$ we write $\Delta_{t,s} = t - s > 0$ for its span.

For online RL, denoising is cast as a sequential decision process (Black et al., 2024): for the transition from $t$ to $s$, the state is the prompt $\boldsymbol{c}$ together with the current latent $\boldsymbol{x}_t$, and the next latent

$\boldsymbol{x}_s$ is an action drawn from a stochastic policy $q_\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t, \boldsymbol{c})$. GRPO (Shao et al., 2024; Liu et al., 2025) optimizes $\theta$ without a learned value function. For each prompt, a group of $M$ trajectories is sampled from the current policy $\theta_{\text{old}}$, which is frozen for one iteration, and the terminal rewards are converted into group-relative advantages $A$. The reward-wise normalization and multi-reward combination used in our experiments are specified in Appendix B. With the policy ratio $r_{t,s}(\theta) = q_\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t, \boldsymbol{c})/q_{\theta_{\text{old}}}(\boldsymbol{x}_s \mid \boldsymbol{x}_t, \boldsymbol{c})$, a standard reference-regularized objective is

$$\begin{aligned}\mathcal{L}_{\text{GRPO}} = & -\mathbb{E}\left[\min\left(r_{t,s}(\theta)A,\ \text{clip}\left(r_{t,s}(\theta), 1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}}\right)A\right)\right] \\ & + \lambda_{\text{KL}}\,\mathbb{E}\left[D_{\text{KL}}\left(q_\theta(\cdot \mid \boldsymbol{x}_t, \boldsymbol{c}) \,\|\, q_{\text{ref}}(\cdot \mid \boldsymbol{x}_t, \boldsymbol{c})\right)\right],\end{aligned} \tag{3}$$

where expectations run over prompts and group trajectories, and $q_{\text{ref}}$ is a frozen reference policy, typically the initial model. The sampling policy $\theta_{\text{old}}$ defines the ratio and is refreshed every iteration; $q_{\text{ref}}$ anchors training. Section 3 analyzes the divergence between $q_{\theta_{\text{old}}}$ and $q_\theta$ induced by one update, while the reference KL in Eq. 3 regularizes training.

### 2.2 Stochastic Transitions for Few-Step Flow RL

Flow matching ordinarily generates samples through a deterministic ODE. Substituting Eq. 2, the Euler update from $t$ to $s$ can be rewritten as $\boldsymbol{x}_t - \Delta_{t,s}\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) = (1-s)\hat{\boldsymbol{x}}_0 + s\hat{\boldsymbol{\epsilon}}$ (Appendix A.1). Stochastic transitions used in flow RL inject Gaussian noise into this update, and those considered in this paper can all be written in the common form

$$\boldsymbol{x}_s = (1-s)\hat{\boldsymbol{x}}_0 + \beta_{t,s}\hat{\boldsymbol{\epsilon}} + \alpha_{t,s}\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{4}$$

with $\boldsymbol{\xi}$ sampled independently; $\beta_{t,s}$ scales the predicted noise and $\alpha_{t,s}$ the newly injected noise. Appendix A.1 compares these coefficients in Table 5 and gives the derivations and implementation variants.

**Flow-GRPO.** Flow-GRPO (Liu et al., 2025) constructs a continuous-time SDE whose marginals coincide with those of the flow ODE and discretizes each transition with Euler–Maruyama, using $\sigma_t = a\sqrt{t/(1-t)}$ with a global noise level $a > 0$. For the standard uncapped construction at interior timesteps, the clean-sample coefficient remains $1-s$, while $\alpha_{t,s}^2 + \beta_{t,s}^2 - s^2$ is strictly positive and grows quadratically with the span at fixed $t$ and $\sigma_t$ (Table 5). These are algebraic properties of the transition coefficients; they do not by themselves determine the sample marginals under learned predictions or establish a performance disadvantage. Appendix A.1 treats the schedule endpoint and noise cap.

**Flow-CPS.** Flow-CPS (Wang & Yu, 2025) instead constructs the transition directly at the target noise level, parameterizing the noise split by a global normalized stochasticity parameter $\phi \in (0, 1]$, corresponding to an angle $\pi\phi/2$:

$$\alpha_{t,s} = s\sin(\pi\phi/2), \qquad \beta_{t,s} = s\cos(\pi\phi/2). \tag{5}$$

These coefficients satisfy $\alpha_{t,s}^2 + \beta_{t,s}^2 = s^2$ for every span, with the ODE Euler update at the deterministic boundary $\phi = 0$. The constraint leaves a one-parameter family of $(\alpha_{t,s}, \beta_{t,s})$: every admissible $\phi$ preserves the coefficients, and Flow-CPS selects $\phi$ empirically. Section 3 optimizes this choice.

## 3 Geo-FSRL

Geo-FSRL selects a stochastic training policy by minimizing the conditional KL divergence induced by a prescribed nonzero velocity-output update. We first formulate this criterion, then derive an equivalent scalar optimization and its closed-form solution. Matched weights subsequently normalize explicit policy-gradient and reference-regularization scales without changing the selected policy's likelihood ratios.

### 3.1 Admissible Stochastic Policies

For a transition with $0 < s < t \le 1$, we use the common form in Eq. 4 and impose

$$\alpha_{t,s} > 0, \qquad 0 \le \beta_{t,s} < s, \qquad \alpha_{t,s}^2 + \beta_{t,s}^2 = s^2. \tag{6}$$

Thus $\beta_{t,s}$ indexes the policies, with $\alpha_{t,s} = \sqrt{s^2 - \beta_{t,s}^2}$. Fixing the total noise coefficient prevents trivially reducing conditional KL by injecting unbounded noise. Under oracle reconstruction, the predicted and fresh noise are independent standard Gaussians, reproducing the linear path's noise coefficient at $s$. Our conditional analysis requires only Gaussian fresh noise $\boldsymbol{\xi}$; the learned noise prediction need not be standard Gaussian.

Substituting the reconstruction estimates into the transition gives

$$\begin{aligned} q_\beta(\cdot \mid \boldsymbol{v}) &= \mathcal{N}\big(\boldsymbol{\mu}_\beta(\boldsymbol{v}), \alpha_{t,s}^2 \boldsymbol{I}\big)\,, \\ \boldsymbol{\mu}_\beta(\boldsymbol{v}) &= (1 - s + \beta_{t,s})\boldsymbol{x}_t + \kappa_{t,s}\boldsymbol{v}, \qquad \kappa_{t,s} = -t(1-s) + (1-t)\beta_{t,s}. \end{aligned} \tag{7}$$

Fix a state $(\boldsymbol{x}_t, \boldsymbol{c})$ and timestep pair $(t, s)$. Let $\boldsymbol{v}_0$ be the old velocity prediction and $\Delta\boldsymbol{v} = \boldsymbol{v}_\theta - \boldsymbol{v}_0 \in \mathbb{R}^D$ an output update, where $D$ is the latent dimension. Each candidate $\beta$ keeps its coefficients and covariance fixed between old and updated policies; across candidates, we prescribe the same $\Delta\boldsymbol{v}$ independently of the optimizer.

### 3.2 Policy Selection by Conditional KL

**A common measure of model change.** The velocity output provides a common measure of model change across stochastic constructions. At the fixed state, the deterministic Euler update changes by $-(t-s)\Delta\boldsymbol{v}$, independently of the training noise split. Yet the same $\Delta\boldsymbol{v}$ can change stochastic transition distributions differently. Coefficient selection therefore determines how a model-output change is expressed in its training policy.

We seek to limit this distribution change for a prescribed output update. Within our Gaussian family, smaller conditional KL implies lower conditional log-ratio variance under each candidate's old policy and, under the threshold condition in Section 3.4, a lower probability of leaving the clipping window. Loss reweighting can normalize explicit gradient and reference-penalty scales but leaves likelihood ratios for a prescribed output update unchanged, so these differences persist after scale matching.

**Policy-selection criterion.** For each candidate, we measure the distribution change induced by an output update using the conditional KL divergence between its updated and old policies:

$$K_\beta(\Delta\boldsymbol{v}) = D_{\mathrm{KL}}(q_\beta(\cdot \mid \boldsymbol{v}_0 + \Delta\boldsymbol{v}) \,\|\, q_\beta(\cdot \mid \boldsymbol{v}_0))\,. \tag{8}$$

Our selection principle is to *choose the admissible policy that minimizes the conditional KL divergence induced by the same prescribed nonzero velocity-output update*. At fixed $(\boldsymbol{x}_t, \boldsymbol{c}, t, s)$ and $\Delta\boldsymbol{v} \neq \boldsymbol{0}$, this defines the optimization problem

$$\min_{0 \le \beta_{t,s} < s} K_\beta(\Delta\boldsymbol{v}). \tag{9}$$

The criterion selects a policy for a prescribed update; reward feedback determines which output updates are useful. We next solve this conditional KL problem within the admissible family.

### 3.3 Geometry Factor and Optimal Transition

**A scalar geometry factor.** To analyze Eq. 9, note that an output update shifts the policy mean by $\kappa_{t,s}\Delta\boldsymbol{v}$ while leaving its covariance fixed. We define

$$G_{t,s}(\beta_{t,s}) = \frac{|\kappa_{t,s}|}{\alpha_{t,s}} = \frac{t(1-s) - (1-t)\beta_{t,s}}{\sqrt{s^2 - \beta_{t,s}^2}}. \tag{10}$$

This *geometry factor* is the standardized policy-mean displacement per unit velocity change: $\|\Delta\boldsymbol{\mu}\|_2/\alpha_{t,s} = G_{t,s}\|\Delta\boldsymbol{v}\|_2$. It depends only on the timestep pair and coefficient choice, and is strictly positive on the admissible family.

**Proposition 1 (Equivalence to the conditional KL criterion)** *At a fixed state and timestep pair, every admissible coefficient choice and every finite output update satisfy*

$$K_\beta(\Delta\boldsymbol{v}) = \frac{1}{2} G_{t,s}^2 \|\Delta\boldsymbol{v}\|_2^2. \tag{11}$$

*For any prescribed $\Delta \boldsymbol{v} \neq \mathbf{0}$, the selection problem in Eq. 9 is therefore equivalent to minimizing $G_{t,s}$:*

$$\arg\min_{0\leq\beta_{t,s}<s} K_\beta(\Delta\boldsymbol{v}) = \arg\min_{0\leq\beta_{t,s}<s} G_{t,s}^2(\beta_{t,s}) = \arg\min_{0\leq\beta_{t,s}<s} G_{t,s}(\beta_{t,s}). \tag{12}$$

*In particular, its minimizing coefficients are independent of the prescribed update's nonzero magnitude and direction.*

*Proof.* Within each candidate, the old and updated policies have equal covariance $\alpha_{t,s}^2 \boldsymbol{I}$ and mean difference $\kappa_{t,s}\Delta\boldsymbol{v}$. The Gaussian KL formula gives $K_\beta = \|\kappa_{t,s}\Delta\boldsymbol{v}\|_2^2/(2\alpha_{t,s}^2)$, which is Eq. 11. Since $\|\Delta\boldsymbol{v}\|_2^2/2$ is a positive constant across candidates and $G_{t,s} > 0$, the three sets of minimizers coincide. □

The equivalence is exact for finite updates; its direction independence follows from isotropic covariance and scalar velocity coupling (Appendix A.2).

**Closed-form solution.**

**Theorem 1 (Geometry-optimal transition)** *For every $0 < s < t \leq 1$ and prescribed $\Delta\boldsymbol{v} \neq \mathbf{0}$, the selection problem in Eq. 9 has the unique solution*

$$\beta_{t,s}^\star = \frac{s^2(1-t)}{t(1-s)}, \qquad \alpha_{t,s}^\star = \sqrt{s^2 - (\beta_{t,s}^\star)^2}, \qquad G_{t,s}^\star = \frac{\sqrt{(t-s)(t+s-2ts)}}{s}. \tag{13}$$

See Appendix A.3 for the proof. In Flow-CPS's parameterization, $\cos(\pi\phi_{t,s}^\star/2) = s(1-t)/(t(1-s))$, so each timestep pair determines its own noise split. Within the admissible family, this rule replaces empirical selection of a global $\phi$ with transition-wise coefficients computed directly from $(t, s)$. At $t = 1$, $\beta_{1,s}^\star = 0$ and $\alpha_{1,s}^\star = s$, allocating all target noise to fresh noise. As $t \downarrow s$ at fixed $s \in (0, 1)$, $\beta_{t,s}^\star \to s$ and $\alpha_{t,s}^\star \to 0$. For $t < 1$, the optimizer balances mean sensitivity and injected noise: $\alpha_{t,s}^\star < s$, whereas maximum noise occurs at $\beta = 0$. Since $G_{t,s}(0)^2 - (G_{t,s}^\star)^2 = (1-t)^2$, these choices coincide only at $t = 1$. The coefficients recover the one-step SDE-DPM-Solver++ transition (Lu et al., 2022), equivalently the posterior-variance ancestral transition under the matching marginal parameterization (Appendix A.3.3), now characterized as optimal for the stated criterion.

**Policy sensitivity and likelihood ratios.** For any nonzero $\Delta\boldsymbol{v}$,

$$\frac{2K_\beta(\Delta\boldsymbol{v})}{\|\Delta\boldsymbol{v}\|_2^2} = G_{t,s}^2, \qquad \mathcal{I}_{\boldsymbol{v}} = G_{t,s}^2 \boldsymbol{I}, \tag{14}$$

where $\mathcal{I}_{\boldsymbol{v}}$ is the conditional Fisher information with respect to the velocity output (Appendix A.2). Here $G_{t,s}^2$ quantifies *policy sensitivity*: twice the conditional KL per squared velocity-update norm.

GRPO evaluates output changes through transition likelihood ratios. For a prescribed $\Delta\boldsymbol{v}$, consider $r = q_\beta(\boldsymbol{x}_s \mid \boldsymbol{v}_0 + \Delta\boldsymbol{v})/q_\beta(\boldsymbol{x}_s \mid \boldsymbol{v}_0)$ with $\boldsymbol{x}_s \sim q_\beta(\cdot \mid \boldsymbol{v}_0)$. Its conditional log-ratio variance satisfies

$$\mathrm{Var}_{q_\beta(\cdot|\boldsymbol{v}_0)}(\log r) = G_{t,s}^2\|\Delta\boldsymbol{v}\|_2^2 = 2K_\beta(\Delta\boldsymbol{v}). \tag{15}$$

Thus the same coefficient choice minimizes both the conditional KL and the log-ratio variance for the same output update, with actions sampled from each candidate's old policy. Appendix A.2.2 derives the full log-ratio distribution and ratio moments. Section 3.4 connects this log-ratio characterization to the dimension-averaged ratio and clipping window used in training.

**Dependence on the transition span.** Even after optimization, $G_{t,s}^\star$ increases strictly with $t$ at fixed $s \in (0, 1)$, from order $\sqrt{t-s}$ for small spans to $G_{1,s}^\star = (1-s)/s$ at the pure-noise endpoint (Appendix A.3). Larger spans therefore increase the minimum conditional KL and log-ratio variance induced by the same velocity update under this fixed-target comparison.

### 3.4 Geometry-Matched GRPO Training

**Transition-wise weights.** The geometry factor also enters the explicit score scale and reference-KL curvature. We normalize these using the gradient-scale matching principle of Team et al. (2025).

At a fixed action, the velocity-output score is

$$\nabla_{\boldsymbol{v}} \log q_\theta = \frac{\kappa^\star_{t,s}}{\alpha^\star_{t,s}} \boldsymbol{\xi}_\theta, \qquad \boldsymbol{\xi}_\theta = \frac{\boldsymbol{x}_s - \boldsymbol{\mu}_{\theta,t,s}}{\alpha^\star_{t,s}}. \tag{16}$$

The explicit coefficient has magnitude $G^\star_{t,s}$. The current and reference policies use the same coefficients and covariance, so their conditional KL has velocity-space curvature $(G^\star_{t,s})^2 \boldsymbol{I}$. These factors vary across timestep pairs even after coefficient selection. We therefore use

$$w^{\text{policy}}_{t,s} = \frac{c_{\text{geo}}}{G^\star_{t,s}}, \qquad w^{\text{KL}}_{t,s} = \left(\frac{c_{\text{geo}}}{G^\star_{t,s}}\right)^2, \tag{17}$$

where $c_{\text{geo}} > 0$ is shared across transitions and the weights are constant with respect to model parameters. The derivation is given in Appendix A.4.

**Training objective.** Following Flow-GRPO (Liu et al., 2025), we average log-densities over the $D$ scalar elements of the transition latent, i.e., all non-batch dimensions. Thus the clipped term uses $\bar{r}_{t,s} = r_{t,s}^{1/D}$ and the reference regularizer uses $\overline{D}_{\text{KL}} = D_{\text{KL}}/D$:

$$\mathcal{L}_{\text{geo}} = \mathbb{E}\Big[ w^{\text{policy}}_{t,s} \bar{\ell}^{\text{clip}}_{t,s}(\theta) + \lambda_{\text{KL}} w^{\text{KL}}_{t,s} \overline{D}_{\text{KL}}(q_\theta \,\|\, q_{\text{ref}}) \Big]. \tag{18}$$

Here $\bar{\ell}^{\text{clip}}$ is the negative clipped surrogate in Eq. 3, with $r$ replaced by $\bar{r}$, and both policies are conditioned on the recorded state. At fixed $D$, dimension averaging changes the KL scale but leaves the minimizing coefficients and the relative $1/G^\star$ and $1/(G^\star)^2$ weights unchanged. Advantage processing and hyperparameters are specified in Appendix B.

**Conditional KL after scale matching.** For any candidate policy, applying its matched weight to the dimension-averaged velocity score at $\theta = \theta_{\text{old}}$ gives $-c_{\text{geo}}\boldsymbol{\xi}_{\text{old}}/D$. The weighted reference KL is $c^2_{\text{geo}}\|\boldsymbol{v}_\theta - \boldsymbol{v}_{\text{ref}}\|_2^2/(2D)$, so its penalty for a fixed velocity deviation is also identical across matched candidates. These cancellations normalize explicit loss scales. They do not change each candidate's sampling distribution, likelihood ratio, or clipping thresholds.

For a prescribed output update at a fixed state, timestep pair, and dimension, let $d = G_{t,s}\|\Delta\boldsymbol{v}\|_2$. With an action drawn from the candidate's old policy, the dimension-averaged log-ratio satisfies

$$\log \bar{r} \sim \mathcal{N}\left( -\frac{d^2}{2D}, \frac{d^2}{D^2} \right). \tag{19}$$

Thus minimizing $G_{t,s}$ still minimizes the conditional KL and log-ratio variance induced by the same output update after matching. It also minimizes the probability that $\bar{r}$ lies outside $[1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}}]$ whenever $-\log(1-\varepsilon_{\text{clip}}) \le 3\log(1+\varepsilon_{\text{clip}})$, as holds for our $\varepsilon_{\text{clip}} = 10^{-3}$ (Proposition 2, Appendix A.2.4). This gives a direct connection between coefficient selection and the clipping window used in training.

This probability compares prescribed updates under the respective old-policy distributions. Actual updates also depend on sampled advantages, residuals, model Jacobians, and the optimizer, while active clipping additionally depends on the advantage sign. Scale matching therefore does not guarantee identical realized updates or a lower observed clipping frequency. For a recorded action reused after an update, $\boldsymbol{\xi}_\theta = \boldsymbol{\xi}_{\text{old}} - (\kappa_{t,s}/\alpha_{t,s})\Delta\boldsymbol{v}$; the resulting dependence of later gradients on the coefficients is detailed in Appendix A.4.

**Effect on the sampling distribution.** At the same state and velocity prediction, changing the coefficients generally changes both the conditional mean and fresh-noise variance: $\boldsymbol{\mu}_{\beta_2} - \boldsymbol{\mu}_{\beta_1} = (\beta_2 - \beta_1)\hat{\boldsymbol{\epsilon}}$. A global learning-rate or clipping adjustment leaves this sampling distribution unchanged. Moreover, geometry-factor ratios generally vary across transitions. Coefficient selection therefore changes the policy beyond a uniform rescaling of its optimization steps.

The sampling and training procedure is given in Algorithm 1 (Appendix B.2).

# 4 EXPERIMENTS

We evaluate four-step image and video generation with Geo-FSRL, Flow-GRPO (Liu et al., 2025), Flow-CPS (Wang & Yu, 2025), and a four-step DiffusionNFT adaptation (Zheng et al., 2026). GRPO-based comparisons share single-stochastic-transition rollouts. Flow-CPS and Geo-FSRL use gradient-scale matching with weights from their respective coefficients unless ablated. Both tasks use 4 sampling steps in training and evaluation. For each task, all reward curves and tabulated automatic scores use the same fixed validation set of 256 prompts, disjoint from training prompts. The horizontal axes count optimizer updates. Training uses 32 NVIDIA H100 GPUs; hyperparameters, schedules, and reward normalization appear in Appendix B.

## 4.1 FEW-STEP TEXT-TO-IMAGE

**Experimental Setup**. We fine-tune all diffusion-transformer parameters of Qwen-Image-2512-Lightning-4steps (Wu et al., 2025; Contributors, 2025), distilled from the 20B MMDiT Qwen-Image, at $720 \times 720$ resolution.

**Reward Models**. We optimize HPSv3 (Ma et al., 2025) for human preference, alone or jointly with CLIPScore (Radford et al., 2021) for image-text alignment.

**Dataset**. Prompts come from the text-to-image generation split of OpenGPT-4o-Image (Chen et al., 2025), a diverse corpus of about 40k prompts.

**Single-reward Training Results**. We compare Geo-FSRL with Flow-CPS at normalized stochasticity $\phi \in \{0.3, 0.5, 0.7, 1.0\}$, using each policy's matched weights. Geo-FSRL derives coefficients from each timestep pair. It improves similarly to the best tested Flow-CPS setting, $\phi = 0.7$, in early and middle training (Figure 1(a)), but finishes higher: 14.05 versus 13.90; Flow-GRPO reaches 13.62 (Table 1). Maximum injected noise, $\phi = 1$, is included and differs from the geometry optimum at internal timesteps. The tested DiffusionNFT adaptation exhibits early reward deterioration and visible pixel noise; this observation is specific to that adaptation.

A global Flow-CPS $\phi$ specifies one geometry factor per stochastic transition. Changing $\phi$ can improve the factor at one transition while worsening it at another, whereas the closed-form schedule minimizes each separately. The scan therefore compares globally fixed noise splits with a transition-wise rule, rather than testing a monotonic relation between any single geometry factor and final reward.

**Multi-reward Joint Training Results**. Joint HPSv3 and CLIPScore training retains the best single-reward Flow-CPS setting, $\phi = 0.7$. Geo-FSRL leads both final metrics (Figure 1(b,c), Table 1): 13.80 HPSv3 versus Flow-CPS's 13.58, and 1.008 CLIPScore versus Flow-GRPO's 0.996, absolute gains of 0.22 and 0.012 over the respective strongest baselines. Qualitative examples appear in Appendix D.

**Human evaluation.** Five evaluators blinded to method identity compare Geo-FSRL and Flow-CPS on 20 image pairs from randomly sampled prompts, with shared prompts and seeds and shuffled presentation order. Geo-FSRL receives 54% of Visual Quality votes (aesthetics) and 58% of Overall Preference votes (prompt adherence and visual appeal; Table 2). These percentages are descriptive results from a small evaluation.

Table 1: Final validation reward scores for single-reward setting and multi-reward setting text-to-image training.

| | Single-Reward | Multi-Reward | |
|---|---|---|---|
| Method | HPSv3 | CLIPScore | HPSv3 |
| Qwen-Image-Lightning | 12.17 | 0.971 | 12.17 |
| Flow-GRPO (Liu et al., 2025) | 13.62 (+1.45) | 0.996 (+0.025) | 13.45 (+1.28) |
| Flow-CPS (Wang & Yu, 2025) | 13.90 (+1.73) | 0.988 (+0.017) | 13.58 (+1.41) |
| Geo-FSRL (Ours) | **14.05** (+1.88) | **1.008** (+0.037) | **13.80** (+1.63) |

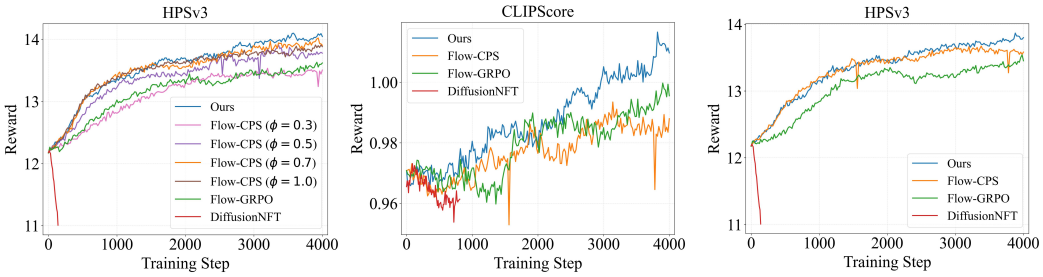


Figure 1: Four-step image validation rewards versus optimizer updates: (a) HPSv3-only; (b,c) joint HPSv3 and CLIPScore. Flow-CPS curves use matched weights for their respective $\phi$ values. DiffusionNFT denotes the tested four-step adaptation.

Table 2: Human evaluation of 20 image pairs by five evaluators blinded to method identity. We report the percentage of preference votes for each method.

| | Visual Quality (%) | Overall Preference (%) |
|---|---|---|
| Flow-CPS | 46 | 42 |
| Geo-FSRL (Ours) | **54** | **58** |

Table 3: Single-reward policy-weighting ablation on the validation set. Each method uses its own matched weights; KL weighting remains enabled throughout. Gain is the change from enabling policy weighting.

| Method | Without policy reweighting | With policy reweighting | Gain |
|---|---|---|---|
| Flow-CPS | **13.69** | 13.90 | +0.21 |
| Geo-FSRL (Ours) | 13.61 | **14.05** | **+0.44** |

**Policy Reweighting Ablation.** Table 3 varies policy weighting while retaining matched KL weighting in all four configurations. Policy weighting improves Flow-CPS by 0.21 and Geo-FSRL by 0.44. Geo-FSRL scores 0.08 below Flow-CPS without policy weighting and 0.15 above with their respective matched weights. The observed ordering therefore changes with policy weighting. These results support the combined design in this setting, evaluating coefficient selection together with its corresponding weights. Appendix C provides the validation curves and a controlled gradient-scale analysis.

### 4.2 Few-Step Text-to-Video

**Experimental Setup**. We fine-tune all diffusion-transformer parameters of LightX2V's four-step distilled Wan2.1-T2V-1.3B (Wan et al., 2025; Contributors, 2025), generating 81 frames at 832 $\times$ 480 resolution.

**Reward Models**. We jointly optimize HPSv3-General and HPSv3-Percentile for frame-level visual quality and caption alignment, and VideoAlign-MQ for motion quality (Liu et al., 2026; Team et al., 2025); Appendix B specifies aggregation.

**Dataset**. Following Self-Forcing (Huang et al., 2026), we use prompts from LLM-extended VidProM (Wang & Yang, 2024).

**Results**. Geo-FSRL leads all three final metrics (Figure 2, Table 4), gaining 1.09 HPSv3-General, 0.80 HPSv3-Percentile, and 0.81 VideoAlign-MQ over the respective strongest baselines. Validation curves show a clearer advantage than for single-reward images. These results evaluate the complete method on a second modality; component comparisons remain in the image ablation. Selected video frames appear in Appendix D.

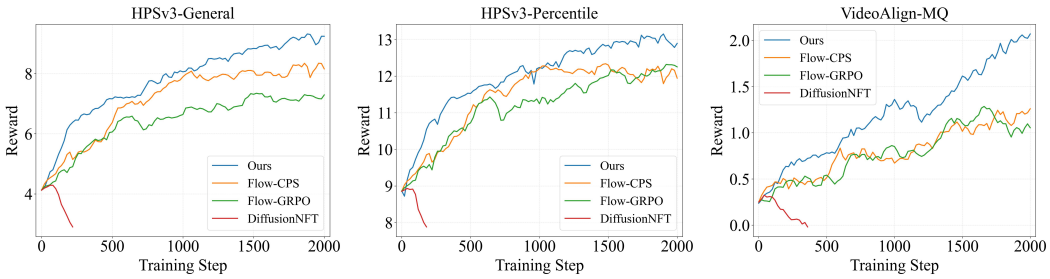


Figure 2: Four-step video validation rewards versus optimizer updates under joint optimization. Flow-CPS uses matched weights; DiffusionNFT denotes the tested four-step adaptation.

Table 4: Final validation reward scores for joint text-to-video training.

| Method | HPSv3-General | HPSv3-Percentile | VideoAlign-MQ |
|---|---|---|---|
| Wan2.1-T2V-1.3B | 4.13 | 8.85 | 0.25 |
| Flow-GRPO (Liu et al., 2025) | 7.29 (+3.16) | 12.10 (+3.25) | 1.05 (+0.80) |
| Flow-CPS (Wang & Yu, 2025) | 8.15 (+4.02) | 11.94 (+3.09) | 1.26 (+1.01) |
| Geo-FSRL (Ours) | **9.24** (+5.11) | **12.90** (+4.05) | **2.07** (+1.82) |

## 5 Related Work

**Policy optimization for fast generation.** DDPO and DPOK use policy gradients for stochastic denoising (Black et al., 2024; Fan et al., 2023). Their reference implementations use DDIM ($\eta = 1$) (Song et al., 2021), yielding ancestral posterior-variance transitions on the timestep grid; their formulations do not require this choice.[1] RLCM applies RL to few-step consistency models with Gaussian re-noising of endpoint predictions (Oertell et al., 2024); our setting retains the velocity parameterization of distilled flows.

**Stochastic policies for flow RL.** Flow-GRPO and DanceGRPO apply group-relative optimization to stochastic flow trajectories (Liu et al., 2025; Xue et al., 2025). Flow-CPS demonstrates four-step training and evaluation with coefficient-preserving transitions (Wang & Yu, 2025). Within this family, our conditional KL criterion (Section 3.2) selects the existing first-order SDE-DPM-Solver++ transition (Lu et al., 2022). Our policy-gradient and KL weights specialize LongCat-Video's gradient-scale matching principle (Team et al., 2025).

**Allocation of stochasticity and computation.** MixGRPO confines stochastic sampling and policy updates to a moving SDE window (Li et al., 2025); TempFlow-GRPO combines designated stochastic branching, noise-aware weighting, and seed grouping (He et al., 2026); BranchGRPO restructures rollouts (Li et al., 2026). Our rollout shares this mixed stochastic–deterministic structure; we address coefficient selection within a stochastic transition.

## 6 Conclusion

We introduce a closed-form, locally best-conditioned transition policy with the corresponding reweighting strategy in this paper. By minimizing the Likelihood-ratio and KL sensitivity and balancing the gradient scale variation across transitions, our method effectively improves reinforcement learning for few-step flow models. In both few-step text-to-image and few-step text-to-video experiments, our method demonstrates faster and larger reward improvement, outperforming the baseline methods. We believe this work provides an effective approach for improving few-step generation and could inspire further research on reinforcement learning for few-step generative models.

[1] Default settings: DDPO and DPOK.

## AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and literature finding to improve presentation clarity and help retrieve related works. We have not used generative AI tools for research ideation, methodological and experimental design, or result analysis. We take responsibility for the final content of this work.

## REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, et al. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*, 2025.

LightX2V Contributors. Lightx2v: Light video generation inference framework. `https://github.com/ModelTC/lightx2v`, 2025.

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.

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. In *International Conference on Learning Representations*, volume 2026, pp. 132869–132895, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, et al. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Yuming Li, Yikai Wang, Zhongyu Zhao, Ming Lu, Qi She, Shanghang Zhang, et al. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 20839–20866, 2026.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 40783–40818. Curran Associates, Inc., 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

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

Owen Oertell, Jonathan D. Chang, Yiyi Zhang, Kianté Brantley, and Wen Sun. RL for consistency models: Faster reward guided text-to-image generation, 2024. URL `https://arxiv.org/abs/2404.03673`.

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.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report, 2025. URL https://arxiv.org/abs/2510.22200.

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.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

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.

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.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

# A Additional Derivations and Proofs

## A.1 Coefficient Form of Finite Flow Transitions

Table 5: Finite transitions from $t$ to $s$ ($0 < s < t < 1$) in the form of Eq. 4. The last column is the deviation from the coefficient-preserving condition $\alpha_{t,s}^2 + \beta_{t,s}^2 = s^2$. The Flow-GRPO row is the standard uncapped Euler–Maruyama construction at interior timesteps; endpoint and noise-cap variants are given in Appendix A.1.

| Transition | $\alpha_{t,s}$ | $\beta_{t,s}$ | $\alpha_{t,s}^2 + \beta_{t,s}^2 - s^2$ |
|---|---|---|---|
| ODE Euler | $0$ | $s$ | $0$ |
| Flow-GRPO (Euler–Maruyama) | $\sigma_t\sqrt{\Delta_{t,s}}$ | $s - \dfrac{\Delta_{t,s}\sigma_t^2}{2t}$ | $\dfrac{\Delta_{t,s}^2\sigma_t^2}{t}\left(1 + \dfrac{\sigma_t^2}{4t}\right)$ |
| Flow-CPS | $s\sin(\pi\phi/2)$ | $s\cos(\pi\phi/2)$ | $0$ |

This section derives the unified coefficient representation of Eq. 4 for the entries of Table 5, with emphasis on the standard Euler–Maruyama transition of Flow-GRPO and the modifications used in its implementation at the schedule endpoint and under a capped noise scale. The coefficients of Flow-CPS and of the policies studied in this paper are given directly in the main text. Throughout, $0 < s < t < 1$ unless stated otherwise, $\Delta_{t,s} = t - s$, and the estimates $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_t - t\boldsymbol{v}_\theta$ and $\hat{\boldsymbol{\epsilon}} = \boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta$ are evaluated at $(\boldsymbol{x}_t, t)$; the arguments of $\boldsymbol{v}_\theta$ are omitted.

### A.1.1 Deterministic Euler Step

Inverting the definitions of the estimates gives $\boldsymbol{x}_t = (1-t)\hat{\boldsymbol{x}}_0 + t\hat{\boldsymbol{\epsilon}}$ and $\boldsymbol{v}_\theta = \hat{\boldsymbol{\epsilon}} - \hat{\boldsymbol{x}}_0$. Hence,

$$\boldsymbol{x}_t - \Delta_{t,s}\boldsymbol{v}_\theta = (1 - t + \Delta_{t,s})\hat{\boldsymbol{x}}_0 + (t - \Delta_{t,s})\hat{\boldsymbol{\epsilon}} = (1-s)\hat{\boldsymbol{x}}_0 + s\hat{\boldsymbol{\epsilon}}, \tag{20}$$

which is the identity quoted in Section 2.2 and the ODE Euler row of Table 5. The identity holds for any velocity output and any span; it rewrites the Euler update in the coefficient form and makes no claim about the accuracy of the update as an ODE solution.

### A.1.2 Flow-GRPO Transition

Flow-GRPO (Liu et al., 2025) samples from the SDE

$$\mathrm{d}\boldsymbol{x}_t = \left[\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) + \frac{\sigma_t^2}{2t}\big(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)\big)\right]\mathrm{d}t + \sigma_t\,\mathrm{d}\boldsymbol{w}_t, \qquad \sigma_t = a\sqrt{\frac{t}{1-t}}, \tag{21}$$

integrated from $t = 1$ toward $t = 0$, where $a > 0$ is the noise level. One Euler–Maruyama step from $t$ to $s$ freezes the coefficients at $(\boldsymbol{x}_t, t)$ and uses the increment $\mathrm{d}t = -\Delta_{t,s}$:

$$\boldsymbol{x}_s^{\mathrm{E}} = \boldsymbol{x}_t - \Delta_{t,s}\left[\boldsymbol{v}_\theta + \frac{\sigma_t^2}{2t}\big(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta\big)\right] + \sigma_t\sqrt{\Delta_{t,s}}\,\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{22}$$

This is the standard update on which the implementation is based; the endpoint and noise-cap modifications are described at the end of this section. The bracketed drift contains $\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta = \hat{\boldsymbol{\epsilon}}$, and the first two terms form the deterministic Euler step of Eq. 20. Therefore,

$$\boldsymbol{x}_s^{\mathrm{E}} = (1-s)\hat{\boldsymbol{x}}_0 + \left(s - \frac{\Delta_{t,s}\sigma_t^2}{2t}\right)\hat{\boldsymbol{\epsilon}} + \sigma_t\sqrt{\Delta_{t,s}}\,\boldsymbol{\xi}, \tag{23}$$

which gives the Flow-GRPO row of Table 5,

$$\alpha_{t,s}^{\mathrm{E}} = \sigma_t\sqrt{\Delta_{t,s}}, \qquad \beta_{t,s}^{\mathrm{E}} = s - \frac{\Delta_{t,s}\sigma_t^2}{2t}. \tag{24}$$

The SDE correction leaves the clean-sample coefficient untouched, shrinks the predicted-noise coefficient by $\Delta_{t,s}\sigma_t^2/(2t)$, and adds fresh noise with standard deviation $\sigma_t\sqrt{\Delta_{t,s}}$.

Expanding the squared sum of the noise coefficients,

$$\begin{aligned}\left(\alpha^{\mathrm{E}}_{t,s}\right)^2+\left(\beta^{\mathrm{E}}_{t,s}\right)^2-s^2 &= \sigma_t^2\Delta_{t,s}-\frac{s\Delta_{t,s}\sigma_t^2}{t}+\frac{\Delta_{t,s}^2\sigma_t^4}{4t^2}\\ &= \frac{\sigma_t^2\Delta_{t,s}(t-s)}{t}+\frac{\Delta_{t,s}^2\sigma_t^4}{4t^2}\\ &= \frac{\Delta_{t,s}^2\sigma_t^2}{t}\left(1+\frac{\sigma_t^2}{4t}\right),\end{aligned} \tag{25}$$

which is the last column of the Flow-GRPO row of Table 5. Both factors are positive, so the squared sum always exceeds $s^2$. Substituting $\sigma_t^2 = a^2 t/(1-t)$ gives the equivalent form

$$\left(\alpha^{\mathrm{E}}_{t,s}\right)^2+\left(\beta^{\mathrm{E}}_{t,s}\right)^2-s^2 = \frac{\Delta_{t,s}^2 a^2}{1-t}\left(1+\frac{a^2}{4(1-t)}\right). \tag{26}$$

**Sign of the predicted-noise coefficient.** From Eq. 23,

$$\beta^{\mathrm{E}}_{t,s}<0 \quad\Longleftrightarrow\quad \Delta_{t,s}\sigma_t^2>2ts \quad\Longleftrightarrow\quad \Delta_{t,s}a^2>2s(1-t). \tag{27}$$

When this holds, the coefficient of the predicted noise changes sign. The condition is easier to satisfy for large spans and for source timesteps close to 1.

**Implementation modifications.** The implementation of the Flow-GRPO transition differs from Eq. 22 only in the noise scale, which is replaced by an effective scale $\tilde{\sigma}_{t,s}$ that accounts for the schedule endpoint and for an optional cap on the injected noise. Let

$$\sigma^{\mathrm{base}}_{t,s}=\begin{cases} a\sqrt{\dfrac{t}{1-t}}, & t<1,\\[2mm] \dfrac{a}{\sqrt{1-s}}, & t=1,\end{cases} \qquad \tilde{\sigma}_{t,s}=\min\left\{\sigma^{\mathrm{base}}_{t,s},\ \frac{c_{\mathrm{clip}}}{\sqrt{\Delta_{t,s}}}\right\}, \tag{28}$$

where $c_{\mathrm{clip}}>0$ is the cap on the standard deviation of the injected noise, and $c_{\mathrm{clip}}=\infty$ when no cap is used. The implemented update is Eq. 22 with $\sigma_t$ replaced by $\tilde{\sigma}_{t,s}$, so Eqs. equation 24–equation 25 hold verbatim with the same substitution, and the injected noise has standard deviation $\alpha^{\mathrm{E}}_{t,s}=\tilde{\sigma}_{t,s}\sqrt{\Delta_{t,s}}\le c_{\mathrm{clip}}$. Two cases are worth recording explicitly.

*First transition before capping.* The base scale $a\sqrt{t/(1-t)}$ diverges as $t\to 1$, so the interior formulas apply to $t<1$ only; at $t_0=1$ the implementation replaces the source timestep by the target timestep inside the denominator $1-t$, i.e. it uses $1-s$ in place of $1-t$, which gives the finite value $\sigma^{\mathrm{base}}_{1,s}=a/\sqrt{1-s}$ for $s<1$. If the cap is inactive, substituting $t=1$ and this scale into Eq. 24 yields

$$\alpha^{\mathrm{E}}_{1,s}=a, \qquad \beta^{\mathrm{E}}_{1,s}=s-\frac{a^2}{2}, \qquad \left(\alpha^{\mathrm{E}}_{1,s}\right)^2+\left(\beta^{\mathrm{E}}_{1,s}\right)^2-s^2=a^2(1-s)+\frac{a^4}{4}, \tag{29}$$

so the first transition injects noise with standard deviation $a$ irrespective of the span. With the cap, the cap threshold at $t=1$ is $c_{\mathrm{clip}}/\sqrt{1-s}$, which has the same denominator as the base scale, so $\alpha^{\mathrm{E}}_{1,s}=\min\{a,c_{\mathrm{clip}}\}$.

*Active cap.* In the region where the cap is active, i.e. $(\sigma^{\mathrm{base}}_{t,s})^2\Delta_{t,s}>c_{\mathrm{clip}}^2$, the effective scale satisfies $\tilde{\sigma}^2_{t,s}=c^2_{\mathrm{clip}}/\Delta_{t,s}$ and Eqs. equation 24–equation 25 become

$$\alpha^{\mathrm{E}}_{t,s}=c_{\mathrm{clip}}, \qquad \beta^{\mathrm{E}}_{t,s}=s-\frac{c^2_{\mathrm{clip}}}{2t}, \qquad \left(\alpha^{\mathrm{E}}_{t,s}\right)^2+\left(\beta^{\mathrm{E}}_{t,s}\right)^2-s^2=\frac{c^2_{\mathrm{clip}}\Delta_{t,s}}{t}+\frac{c^4_{\mathrm{clip}}}{4t^2}. \tag{30}$$

At fixed $t$ and $c_{\mathrm{clip}}$, the deviation is therefore affine in the span inside the active region, and it remains strictly positive, so the capped transition still lies outside the coefficient-preserving family. Outside the active region the uncapped coefficients apply. Whether the cap is active for a given transition depends on the schedule, the noise level, and $c_{\mathrm{clip}}$; the values used in our experiments are reported in Appendix B.

**Terminal boundary.** At $s = 0$ the coefficient-preserving condition reduces to $\alpha_{t,0} = \beta_{t,0} = 0$, and relative deviations normalized by $s^2$ are undefined; this is consistent with keeping the final transition deterministic.

### A.2 Conditional KL and Likelihood-Ratio Geometry

#### A.2.1 KL Equivalence and Fisher Information

**Proof of Proposition 1.** Fix $(\boldsymbol{x}_t, \boldsymbol{c}, t, s)$ and one admissible coefficient choice. The old and updated policies have common covariance $\alpha_{t,s}^2 \boldsymbol{I}$, and their mean difference is

$$\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_\beta(\boldsymbol{v}_0 + \Delta\boldsymbol{v}) - \boldsymbol{\mu}_\beta(\boldsymbol{v}_0) = \kappa_{t,s}\Delta\boldsymbol{v}. \tag{31}$$

The equal-covariance Gaussian KL formula therefore gives, in either direction,

$$K_\beta(\Delta\boldsymbol{v}) = \frac{\|\Delta\boldsymbol{\mu}\|_2^2}{2\alpha_{t,s}^2} = \frac{1}{2}G_{t,s}^2\|\Delta\boldsymbol{v}\|_2^2. \tag{32}$$

This identity is exact for finite output differences. Moreover, $t(1-s) - (1-t)\beta_{t,s} \geq t(1-s) - (1-t)s = t - s > 0$, so $G_{t,s} > 0$ throughout the family. For any prescribed nonzero update, $\|\Delta\boldsymbol{v}\|_2^2/2$ is a positive constant across candidates. Minimizing the conditional KL in Eq. 9 is therefore equivalent to minimizing $G_{t,s}^2$ or $G_{t,s}$, which establishes Eq. 12. For two candidates $A, B$, Eq. 32 also implies $G_A \leq G_B$ if and only if $K_A(\Delta\boldsymbol{v}) \leq K_B(\Delta\boldsymbol{v})$ for every output update; the reverse implication follows by evaluating any nonzero update. The minimizing coefficients are consequently independent of the prescribed update's nonzero magnitude and direction. □

**Normalized KL response and Fisher information.** Dividing Eq. 32 by $\|\Delta\boldsymbol{v}\|_2^2/2$ for a nonzero update gives the normalized KL response $G_{t,s}^2$, interpreted as policy sensitivity in Eq. 14. To obtain the Fisher information, differentiate with the action $\boldsymbol{x}_s$ held fixed:

$$\nabla_{\boldsymbol{v}} \log q_\beta(\boldsymbol{x}_s \mid \boldsymbol{v}) = \frac{\kappa_{t,s}}{\alpha_{t,s}^2}(\boldsymbol{x}_s - \boldsymbol{\mu}_\beta(\boldsymbol{v})). \tag{33}$$

Under the policy at $\boldsymbol{v}$, write $\boldsymbol{x}_s - \boldsymbol{\mu}_\beta(\boldsymbol{v}) = \alpha_{t,s}\boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Hence

$$\mathcal{I}_{\boldsymbol{v}} = \mathbb{E}\big[(\nabla_{\boldsymbol{v}} \log q_\beta)(\nabla_{\boldsymbol{v}} \log q_\beta)^\top\big] = \frac{\kappa_{t,s}^2}{\alpha_{t,s}^2}\boldsymbol{I} = G_{t,s}^2\boldsymbol{I}. \tag{34}$$

#### A.2.2 Likelihood-Ratio Moments

Keep the state and timestep pair fixed, and prescribe an output update $\Delta\boldsymbol{v} \in \mathbb{R}^D$ independently of the action used to evaluate the ratio. For an action from the old policy, write $\boldsymbol{x}_s = \boldsymbol{\mu}_\beta(\boldsymbol{v}_0) + \alpha_{t,s}\boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and set $\boldsymbol{z} = \kappa_{t,s}\Delta\boldsymbol{v}/\alpha_{t,s}$. For $r = q_\beta(\boldsymbol{x}_s \mid \boldsymbol{v}_0 + \Delta\boldsymbol{v})/q_\beta(\boldsymbol{x}_s \mid \boldsymbol{v}_0)$, the Gaussian log-density difference is

$$\begin{aligned} \log r &= -\frac{\|\alpha_{t,s}\boldsymbol{\xi} - \kappa_{t,s}\Delta\boldsymbol{v}\|_2^2 - \|\alpha_{t,s}\boldsymbol{\xi}\|_2^2}{2\alpha_{t,s}^2} \\ &= \boldsymbol{\xi}^\top\boldsymbol{z} - \frac{1}{2}\|\boldsymbol{z}\|_2^2, \qquad \|\boldsymbol{z}\|_2 = G_{t,s}\|\Delta\boldsymbol{v}\|_2 = d. \end{aligned} \tag{35}$$

Since $\boldsymbol{\xi}^\top\boldsymbol{z} \sim \mathcal{N}(0, d^2)$,

$$\log r \sim \mathcal{N}(-d^2/2, d^2), \qquad \mathrm{Var}(\log r) = 2K_\beta(\Delta\boldsymbol{v}). \tag{36}$$

In particular, $-\mathbb{E}_{q_{\mathrm{old}}}[\log r] = d^2/2 = D_{\mathrm{KL}}(q_{\mathrm{old}}\|q_\theta) = K_\beta(\Delta\boldsymbol{v})$, where the last equality uses the common covariance. The log-normal moments are

$$\mathbb{E}_{q_{\mathrm{old}}}[r^m] = \exp\left(\frac{m^2 - m}{2}d^2\right), \qquad \mathbb{E}[r] = 1, \qquad \mathrm{Var}(r) = e^{d^2} - 1. \tag{37}$$

These conditional moments characterize likelihood changes without assumptions on rewards or advantages.

### A.2.3 DIMENSION-AVERAGED LOG-DENSITIES

The implemented ratio is $\bar{r} = \exp((\log q_\theta - \log q_{\text{old}})/D) = r^{1/D}$, where $D \geq 1$ is the latent dimension. Under the same fixed-state, prescribed-update comparison, Eq. 36 gives

$$\log \bar{r} \sim \mathcal{N}\left(-\frac{d^2}{2D}, \frac{d^2}{D^2}\right), \qquad \overline{K}_\beta = \frac{K_\beta}{D} = \frac{G_{t,s}^2}{2D}\|\Delta \boldsymbol{v}\|_2^2. \tag{38}$$

Consequently, $\mathrm{Var}(\log \bar{r}) = 2\overline{K}_\beta/D$. With $D$ fixed across candidate coefficients, minimizing $G_{t,s}$ minimizes both the normalized KL and the averaged log-ratio variance for the same output update.

The quantity $\bar{r}$ is a scaled surrogate, not generally the density ratio of the two policies. Its moments are

$$\mathbb{E}[\bar{r}] = \exp\left(-\frac{d^2}{2D} + \frac{d^2}{2D^2}\right), \qquad \mathrm{Var}(\bar{r}) = \exp\left(-\frac{d^2}{D} + \frac{d^2}{D^2}\right)\left(\exp\left(\frac{d^2}{D^2}\right) - 1\right). \tag{39}$$

For $D > 2$ this variance is not globally monotone in $d^2$ and tends to zero as $d^2 \to \infty$. The KL and log-ratio variance ordering therefore does not imply a global ordering of $\mathrm{Var}(\bar{r})$.

### A.2.4 CLIPPING-WINDOW EXCEEDANCE

**Proposition 2 (Population clipping-window exceedance)** *Fix the state, timestep pair, and dimension $D \geq 1$, and prescribe $\Delta \boldsymbol{v}$ independently of the evaluation draw from each candidate's old policy. Let $0 < \epsilon < 1$, $d = G_{t,s}\|\Delta \boldsymbol{v}\|_2$, and*

$$a = -D\log(1-\epsilon), \qquad b = D\log(1+\epsilon). \tag{40}$$

*For $d > 0$, the probability of leaving the clipping window is*

$$\begin{aligned} p_{\text{out}}(d) &:= \Pr(\bar{r} \notin [1-\epsilon, 1+\epsilon]) \\ &= \Phi\left(\frac{d}{2} - \frac{a}{d}\right) + \Phi\left(-\frac{d}{2} - \frac{b}{d}\right), \end{aligned} \tag{41}$$

*where $\Phi$ is the standard normal distribution function, and $p_{\text{out}}(0) = 0$. If $a \leq 3b$, then $p_{\text{out}}(d)$ is strictly increasing for $d > 0$. Hence, under this condition, minimizing $G_{t,s}$ also minimizes the population clipping-window exceedance probability for every prescribed nonzero output update.*

**Proof.** By Eq. 38, $\log \bar{r} = (-d^2/2 + dZ)/D$ with $Z \sim \mathcal{N}(0,1)$. The lower-window event is $Z < d/2 - a/d$, and the upper-window event is $Z > d/2 + b/d$, which gives Eq. 41. At $d = 0$, $\bar{r} = 1$ almost surely.

Let $\varphi$ denote the standard normal density. Differentiating for $d > 0$ gives

$$p'_{\text{out}}(d) = \varphi\left(\frac{d}{2} - \frac{a}{d}\right)\left(\frac{1}{2} + \frac{a}{d^2}\right) + \varphi\left(\frac{d}{2} + \frac{b}{d}\right)\left(-\frac{1}{2} + \frac{b}{d^2}\right). \tag{42}$$

If $d^2 \leq 2b$, the first term is positive and the second is nonnegative. If $d^2 > 2b$, use $a > b > 0$ and $a \leq 3b$ to obtain

$$\frac{\varphi(d/2 - a/d)}{\varphi(d/2 + b/d)} = \exp\left(\frac{a+b}{2} - \frac{a^2 - b^2}{2d^2}\right) \geq \exp\left(\frac{(a+b)(3b-a)}{4b}\right) \geq 1. \tag{43}$$

Substitution into Eq. 42 then yields $p'_{\text{out}}(d) \geq \varphi(d/2 + b/d)(a+b)/d^2 > 0$. This proves strict monotonicity. □

The condition $a \leq 3b$ is equivalent to $-\log(1-\epsilon) \leq 3\log(1+\epsilon)$ and holds for the experimental choice $\epsilon = 0.001$, independently of $D$. The proposition orders the total population probability of leaving the window. Active PPO clipping additionally depends on the advantage sign: the clipped branch is active above the upper threshold for positive advantages and below the lower threshold for negative advantages. The result also does not establish an ordering of empirical minibatch fractions when the output update depends on the same samples used to evaluate those fractions.

### A.2.5 Fisher Metric in Parameter Space

Let $\boldsymbol{J}_\theta = \partial \boldsymbol{v}_\theta / \partial \theta$ at the fixed conditioning state. The chain rule gives

$$\nabla_\theta \log q_\theta = \boldsymbol{J}_\theta^\top \nabla_{\boldsymbol{v}} \log q_\theta, \qquad \mathcal{I}_\theta = \boldsymbol{J}_\theta^\top \mathcal{I}_{\boldsymbol{v}} \boldsymbol{J}_\theta = G_{t,s}^2 \boldsymbol{J}_\theta^\top \boldsymbol{J}_\theta. \tag{44}$$

For the same model and state, coefficient selection changes the scalar factor multiplying this conditional metric. It does not change the Jacobian or its directional anisotropy. Unlike the exact KL identity in output space, the corresponding quadratic approximation in parameter space uses a local linearization of the model. A shared parameter update need not independently realize prescribed output changes across states.

### A.2.6 Equivalent KL-Budget Interpretation

The same conditional comparison can be expressed as an output-space allowance. For a common budget $\delta > 0$, define $\mathcal{T}_\beta(\delta) = \{\Delta \boldsymbol{v} \in \mathbb{R}^D : K_\beta(\Delta \boldsymbol{v}) \le \delta\}$. Equation 32 gives

$$K_\beta(\Delta \boldsymbol{v}) \le \delta \quad \Longleftrightarrow \quad \|\Delta \boldsymbol{v}\|_2 \le \frac{\sqrt{2\delta}}{G_{t,s}}. \tag{45}$$

Thus the admissible region is an entire centered ball with radius

$$\rho_\beta(\delta) = \frac{\sqrt{2\delta}}{G_{t,s}(\beta_{t,s})}. \tag{46}$$

Comparing these concentric balls yields

$$G_A \le G_B \quad \Longleftrightarrow \quad \mathcal{T}_B(\delta) \subseteq \mathcal{T}_A(\delta) \quad \text{for every } \delta > 0. \tag{47}$$

The same equivalence holds at any one positive budget. Minimizing $G_{t,s}$ therefore also maximizes the radius for every positive budget. For a normalized budget $\overline{K}_\beta \le \bar{\delta}$, the radius is $\sqrt{2D\bar{\delta}}/G_{t,s}$, so the optimum is unchanged at fixed $D$. At zero budget, every candidate permits only the zero update.

## A.3 Geometry-Optimal Transition

### A.3.1 Proof of Theorem 1

Write $A = t(1-s) > 0$ and $B = 1 - t \ge 0$. For $0 \le \beta < s$, $A - B\beta \ge A - Bs = t - s > 0$, so

$$G_{t,s}(\beta) = \frac{A - B\beta}{\sqrt{s^2 - \beta^2}}, \qquad \frac{\mathrm{d}G_{t,s}}{\mathrm{d}\beta} = \frac{A\beta - Bs^2}{(s^2 - \beta^2)^{3/2}}. \tag{48}$$

For $t < 1$, the derivative changes sign from negative to positive at $\beta^\star = Bs^2/A$. This point lies in the domain because $\beta^\star/s = s(1-t)/(t(1-s)) < 1$, which is equivalent to $s < t$. For $t = 1$, $B = 0$ and the derivative is strictly positive for $\beta > 0$, so the unique minimum is the boundary point $\beta^\star = 0$. In both cases $\alpha^\star = \sqrt{s^2 - (\beta^\star)^2}$ is uniquely determined by positivity and the coefficient constraint. Extending the domain to $-s < \beta < s$ leaves the unique minimizer unchanged: $A\beta - Bs^2 < 0$ throughout $\beta < 0$.

Substitution yields

$$(G_{t,s}^\star)^2 = \frac{A^2 - B^2 s^2}{s^2} = \frac{(A - Bs)(A + Bs)}{s^2} = \frac{(t-s)(t+s-2ts)}{s^2}. \tag{49}$$

Taking the positive square root proves the stated minimum. Proposition 1 then establishes that these coefficients minimize the conditional KL for every prescribed nonzero velocity-output update. □

### A.3.2 Span Dependence and Boundaries

The normalized stochasticity parameter satisfies $\cos(\pi \phi_{t,s}^\star/2) = \beta_{t,s}^\star/s = s(1-t)/(t(1-s))$. At fixed $s \in (0,1)$, the derivative of this retained-noise fraction with respect to $t$ is $-s/(t^2(1-s)) < 0$.

The retained noise coefficient fraction therefore decreases as the span grows in this fixed-target comparison. For $t = s + \Delta$,

$$(G^\star_{s+\Delta,s})^2 = \frac{2(1-s)}{s}\Delta + \frac{1-2s}{s^2}\Delta^2, \qquad G^\star_{s+\Delta,s} = \sqrt{\frac{2(1-s)}{s}}\sqrt{\Delta} + O(\Delta^{3/2}). \tag{50}$$

The expansion holds as $\Delta \downarrow 0$ with $s$ fixed in $(0, 1)$. For the full admissible range,

$$\frac{\partial}{\partial t}(G^\star_{t,s})^2 = \frac{2[t(1-2s)+s^2]}{s^2} > 0. \tag{51}$$

Indeed, if $s \le 1/2$, the first term in brackets is nonnegative and $s^2 > 0$. If $s > 1/2$, the bracket is at least $1 - 2s + s^2 = (1-s)^2 > 0$, since $t \le 1$. This proves strict monotonicity. At $t = 1$,

$$\beta^\star_{1,s} = 0, \qquad \alpha^\star_{1,s} = s, \qquad G^\star_{1,s} = \frac{1-s}{s}. \tag{52}$$

The terminal case $s = 0$ is outside the stochastic family: the coefficient constraint forces $\alpha_{t,0} = \beta_{t,0} = 0$, giving the deterministic Euler update and no non-degenerate Gaussian likelihood.

### A.3.3 Correspondence with SDE-DPM-Solver++

For $0 < s < t < 1$, the one-step SDE-DPM-Solver++ transition with data prediction (Lu et al., 2022) is

$$\boldsymbol{x}_s = \frac{\sigma_s}{\sigma_t}e^{-h}\boldsymbol{x}_t + a_s(1 - e^{-2h})\hat{\boldsymbol{x}}_0 + \sigma_s\sqrt{1-e^{-2h}}\boldsymbol{\xi}, \qquad h = \log\frac{a_s/\sigma_s}{a_t/\sigma_t}. \tag{53}$$

Here $a_u$ and $\sigma_u$ denote the clean-sample and noise coefficients of the marginal schedule, distinct from the transition coefficients $\alpha_{t,s}, \beta_{t,s}$. For rectified flow, $a_u = 1 - u$ and $\sigma_u = u$. Setting $b = e^{-h} = s(1-t)/(t(1-s))$ and using $\boldsymbol{x}_t = (1-t)\hat{\boldsymbol{x}}_0 + t\hat{\boldsymbol{\epsilon}}$ gives

$$\begin{aligned} \frac{s}{t}b\boldsymbol{x}_t + (1-s)(1-b^2)\hat{\boldsymbol{x}}_0 &= (1-s)\hat{\boldsymbol{x}}_0 + sb\hat{\boldsymbol{\epsilon}}, \\ sb = \beta^\star_{t,s}, \qquad & s\sqrt{1-b^2} = \alpha^\star_{t,s}. \end{aligned} \tag{54}$$

Thus the coefficients agree exactly for every velocity prediction, without an oracle assumption. The correspondence extends continuously to $t = 1$ through $b = 0$; the log-SNR itself is not finite at that endpoint.

**A Gaussian Markov interpretation.** A related interpretation uses a Gaussian Markov process with the same one-time conditional marginals as the linear probability path. For $0 < s < t \le 1$, define

$$\begin{aligned} \boldsymbol{x}_s \mid \boldsymbol{x}_0 &\sim \mathcal{N}((1-s)\boldsymbol{x}_0, s^2\boldsymbol{I}), \\ \boldsymbol{x}_t \mid \boldsymbol{x}_s &\sim \mathcal{N}(\gamma_{t,s}\boldsymbol{x}_s, (t^2 - \gamma^2_{t,s}s^2)\boldsymbol{I}), \qquad \gamma_{t,s} = \frac{1-t}{1-s}. \end{aligned} \tag{55}$$

The innovation variance is positive because $t > \gamma_{t,s}s$. The resulting marginal is $\boldsymbol{x}_t \mid \boldsymbol{x}_0 \sim \mathcal{N}((1-t)\boldsymbol{x}_0, t^2\boldsymbol{I})$. This Markov coupling differs from interpolating with a single shared noise vector at all times; the marginal path alone does not specify the joint law across time.

Gaussian conditioning gives

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_s \mid \boldsymbol{x}_t, \boldsymbol{x}_0] &= (1-s)\boldsymbol{x}_0 + \frac{s^2\gamma_{t,s}}{t^2}(\boldsymbol{x}_t - (1-t)\boldsymbol{x}_0), \\ \mathrm{Cov}(\boldsymbol{x}_s \mid \boldsymbol{x}_t, \boldsymbol{x}_0) &= \left(s^2 - \frac{s^4\gamma^2_{t,s}}{t^2}\right)\boldsymbol{I}. \end{aligned} \tag{56}$$

Replacing $\boldsymbol{x}_0$ by $\hat{\boldsymbol{x}}_0$ and substituting $\boldsymbol{x}_t - (1-t)\hat{\boldsymbol{x}}_0 = t\hat{\boldsymbol{\epsilon}}$ yields $\beta_{t,s} = s^2\gamma_{t,s}/t = \beta^\star_{t,s}$ and $\alpha_{t,s} = \alpha^\star_{t,s}$. This is a plug-in posterior interpretation of the selected transition, rather than a claim that the learned reconstruction gives the true posterior of the data.

### A.4 Derivation and Scope of Geometry-Matched Scaling

The weights match two explicit scales of the selected conditional policy. For actions sampled from the current policy, Eq. 33 becomes $\nabla_{\boldsymbol{v}} \log q_\theta = \operatorname{sgn}(\kappa^\star_{t,s}) G^\star_{t,s} \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$; the action is held fixed when differentiating. The same identity applies at $\theta = \theta_{\text{old}}$ for recorded rollout actions. Multiplying the score by $w^{\text{policy}} = c_{\text{geo}}/G^\star_{t,s}$ makes this explicit coefficient have magnitude $c_{\text{geo}}$.

For a stored old-policy action evaluated after the model changes, the standardized residual is instead

$$\boldsymbol{\xi}_\theta = \boldsymbol{\xi}_{\text{old}} - \frac{\kappa^\star_{t,s}}{\alpha^\star_{t,s}} \Delta \boldsymbol{v}, \qquad \nabla_{\boldsymbol{v}} \log q_\theta = \frac{\kappa^\star_{t,s}}{\alpha^\star_{t,s}} \boldsymbol{\xi}_{\text{old}} - (G^\star_{t,s})^2 \Delta \boldsymbol{v}. \tag{57}$$

Here $\boldsymbol{\xi}_{\text{old}}$ is standard Gaussian under the old policy. Consequently, canceling the leading coefficient does not remove all geometry dependence of later optimization gradients. Advantages, the likelihood ratio, clipping, the model Jacobian, and the optimizer further affect the parameter update.

For the reference term, both policies use the selected coefficients, so

$$D_{\text{KL}}(q_\theta \| q_{\text{ref}}) = \frac{1}{2} (G^\star_{t,s})^2 \|\boldsymbol{v}_\theta - \boldsymbol{v}_{\text{ref}}\|_2^2, \qquad \nabla^2_{\boldsymbol{v}_\theta} D_{\text{KL}} = (G^\star_{t,s})^2 \boldsymbol{I}. \tag{58}$$

The weight $w^{\text{KL}} = (c_{\text{geo}}/G^\star_{t,s})^2$ maps this curvature to $c^2_{\text{geo}} \boldsymbol{I}$. The same covariance also makes the KL symmetric, although its reference-regularization role differs from the updated-to-old conditional KL used for policy selection.

Dimension averaging introduces a factor $1/D$ into both the log-density score and the reference KL. With the same weights, their matched explicit scales become $c_{\text{geo}}/D$ and $c^2_{\text{geo}}/D$, respectively. Thus the relative transition-wise weights are unchanged at fixed $D$. This does not turn the averaged log-density into the log of a normalized joint policy, or make its score outer product equal to the curvature of the averaged KL.

**Estimator scale and signal-to-noise ratio.** At a fixed state, timestep pair, and latent dimension, with $\theta = \theta_{\text{old}}$, an unclipped negative policy-surrogate gradient has the form $\widehat{\boldsymbol{g}}_\beta = (G_\beta/D)\boldsymbol{Y}_\beta$, where $\boldsymbol{Y}_\beta = A_\beta \boldsymbol{J}^\top \boldsymbol{\xi}$ and $\boldsymbol{J} = \partial \boldsymbol{v}/\partial \theta$. For nonzero variance, the explicit factor cancels from $\|\mathbb{E}\widehat{\boldsymbol{g}}_\beta\|_2 / \sqrt{\operatorname{tr} \operatorname{Cov}(\widehat{\boldsymbol{g}}_\beta)}$. The action–advantage distribution nevertheless depends on the policy, so this cancellation establishes neither equality nor an ordering of signal-to-noise ratios across coefficient choices. For example, in one dimension with $x = \mu + \alpha\xi$, reward $R(x) = x$, and its exact mean as baseline, $A = \alpha\xi$ and the reward-gradient estimator is $\kappa\xi^2$; its signal-to-noise ratio is $1/\sqrt{2}$ for every $\alpha > 0$. Thus increasing noise need not increase estimator signal-to-noise ratio.

**The global matching multiplier.** Writing $P = \mathbb{E}[\bar{\ell}^{\text{clip}}/G^\star]$ and $K = \mathbb{E}[\overline{D}_{\text{KL}}/(G^\star)^2]$ gives $\mathcal{L}(c, \lambda) = cP + \lambda c^2 K$ and $\mathcal{L}(ac, \lambda/a) = a\mathcal{L}(c, \lambda)$ for $a > 0$. Thus $c$ affects the relative reference penalty and an overall loss scale. Plain SGD offsets the latter by dividing the learning rate by $a$. AdamW instead cancels uniform gradient scaling at the same learning rate in the zero-$\epsilon$, unclipped limit with zero-initialized moments. Finite optimizer $\epsilon$ and a fixed gradient-clipping threshold can prevent exact equivalence. The multiplier does not change the sampling policy or its conditional KL cost for a prescribed velocity update.

## B Experimental Details

Table 6 lists the training hyperparameters for the GRPO-based experiments.

Flow-GRPO uses noise level $0.7$ for main image comparisons and $0.3$ for video, and caps the injected transition standard deviation at $0.45$ (Appendix A.1). Flow-CPS uses $\phi = 0.7$ for the main image comparisons and $\phi = 0.3$ for video; the image single-reward scan also tests $0.3$, $0.5$, and $1.0$. Flow-CPS and Geo-FSRL both use $c_{\text{geo}} = 0.5$ in their matched weights. The policy-only ablation sets $w^{\text{policy}} = 1$ and retains each method's $w^{\text{KL}}$.

For each reward, we subtract the prompt-group mean and divide by the largest prompt-group standard deviation in the collected rollout batch, with a stabilizer of $10^{-4}$. In multi-reward training, the resulting reward-specific advantages are summed with equal weights and clipped to $[-5, 5]$.

Table 6: Training hyperparameters for the image and video GRPO-based experiments. The geometry multiplier $c_{\text{geo}}$ applies to the matched weights of Geo-FSRL and Flow-CPS.

| Setting | Text-to-image | Text-to-video |
|---|---|---|
| Optimizer updates | 4,000 | 2,000 |
| Optimizer | AdamW | AdamW |
| Learning rate | $2 \times 10^{-6}$ | $2 \times 10^{-6}$ |
| Adam $(\beta_1, \beta_2)$ | $(0.9, 0.999)$ | $(0.9, 0.999)$ |
| Adam $\epsilon$ | $10^{-8}$ | $10^{-8}$ |
| Weight decay | $10^{-4}$ | $10^{-4}$ |
| Maximum gradient norm | 0.5 | 0.5 |
| Ratio clip range for $\bar{r}$ | $10^{-3}$ | $10^{-3}$ |
| Reference-penalty coefficient | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ |
| Geometry multiplier $c_{\text{geo}}$ | 0.5 | 0.5 |
| Samples per prompt | 8 | 8 |
| Rollout batch size per device | 8 | 4 |
| Rollout batches per iteration | 2 | 2 |
| Training batch size per device | 8 | 4 |
| Optimization passes per rollout iteration | 1 | 1 |
| Gradient accumulation steps | 1 | 1 |
| Advantage clipping | $[-5, 5]$ | $[-5, 5]$ |
| Random seed | 42 | 42 |

**Reported rewards.** For each task, the curves and tabulated automatic reward scores are computed on the same fixed set of 256 validation prompts, shared across methods and held out from training. Curve horizontal axes count optimizer updates. The image CLIP score uses CLIP ViT-L/14 matching logits divided by 30; it is a scaled similarity and need not lie in $[0, 1]$. For video, HPSv3-General averages frame scores using the generic prompt "A high-quality image." HPSv3-Percentile uses the video caption and averages the top 30% of frame scores. All reported automatic metrics are also optimized during training; the human comparison reports descriptive preferences, with shared images and evaluators across ratings.

### B.1 Timestep Schedule and Coefficients

Wan uses $t = 5u/(1 + 4u)$ for $u \in \{1, 3/4, 1/2, 1/4, 0\}$, giving $(1, 0.9375, 0.833333 \ldots, 0.625, 0)$ for both training and evaluation. Qwen uses four steps from the scheduler saved with its Lightning checkpoint. The last transition to $s = 0$ is deterministic and is excluded from policy optimization. Table 7 compares the coefficients on the Wan grid.

Video comparisons use $\phi = 0.3$ for greater geometric separation from the optimum, whose $\phi^\star$ values are approximately $(1, 0.783653, 0.783653)$. With $\phi = 0.7$, $G$ is only $1.03\%$ above $G^\star$ at the two interior-source transitions and $12.23\%$ above at the first. This analytic comparison does not establish a reward ordering between Flow-CPS settings.

Table 7: Analytic Wan coefficients (five decimal places). Flow-CPS with $\phi = 0.3$ is the reported video baseline; $\phi = 0.7$ is shown for analytic comparison.

| | Selected split ($\phi^\star$) | | | Flow-CPS ($\phi = 0.3$) | | | Flow-CPS ($\phi = 0.7$) | | |
|---|---|---|---|---|---|---|---|---|---|
| $(t, s)$ | $\alpha^\star$ | $\beta^\star$ | $G^\star$ | $\alpha$ | $\beta$ | $G$ | $\alpha$ | $\beta$ | $G$ |
| $(1, 15/16)$ | 0.93750 | 0.00000 | 0.06667 | 0.42562 | 0.83532 | 0.14685 | 0.83532 | 0.42562 | 0.07482 |
| $(15/16, 5/6)$ | 0.78567 | 0.27778 | 0.17678 | 0.37833 | 0.74251 | 0.29034 | 0.74251 | 0.37833 | 0.17859 |
| $(5/6, 5/8)$ | 0.58926 | 0.20833 | 0.47140 | 0.28374 | 0.55688 | 0.77424 | 0.55688 | 0.28374 | 0.47624 |

### B.2 Training Algorithm

**Sampling arrangement.** For each rollout batch, we select one non-terminal transition index $k \in \{0, \dots, K-2\}$ and use the stochastic policy only at that step. All other steps use deterministic Euler updates with the frozen sampling model $\theta_{\text{old}}$. Only the selected transition contributes to the loss, with its trajectory's terminal reward determining the advantage; no gradient is propagated through the rollout. The index cycles across rollout batches, and the recorded transitions are shuffled before optimization. The expectation in Eq. 18 is over these rollout batches, prompts, and group trajectories. At inference, the updated velocity model uses the original deterministic few-step solver.

**Algorithm 1** Geo-FSRL Training

**Require:** Velocity model $\theta$; reference $\theta_{\text{ref}}$; schedule $\mathcal{T}$; reward model $\mathcal{R}$; group size $M$; scale $c_{\text{geo}}$
Precompute optimal coefficients and weights for each non-terminal transition using Eqs. 13 and 17
**for** each GRPO iteration **do**
  Freeze $\theta_{\text{old}} \leftarrow \theta$ and initialize $\mathcal{B} \leftarrow \emptyset$
  **for** each rollout batch **do**
    Select $k$ by cycling over $\{0, \dots, K-2\}$
    For each prompt, generate $M$ trajectories with $\theta_{\text{old}}$: sample stochastically at $k$ and use deterministic Euler updates elsewhere
    Evaluate terminal rewards and construct group-relative advantages $A$
    Store $(\boldsymbol{c}, \boldsymbol{x}_{t_k}, \boldsymbol{x}_{t_{k+1}}, k, \log q_{\text{old}}/D, A)$ for each trajectory in $\mathcal{B}$
  **end for**
  **for** each optimization minibatch from shuffled $\mathcal{B}$ **do**
    Evaluate the current log-density and current/reference policy means at the stored states
    Form $\bar{r}$, the dimension-averaged Gaussian reference KL, and the weights for the recorded timestep pairs
    Update $\theta$ by minimizing Eq. 18
  **end for**
**end for**
**Ensure:** Optimized velocity model $\theta$

## C Additional Ablation Results

Figure 3 gives the validation curves for Table 3, with KL weighting retained throughout. The Geo-FSRL–Flow-CPS score difference changes from $-0.08$ without policy weighting to $+0.15$ with matched weighting, an observed interaction contrast of $0.23$. The matched comparison changes both the coefficients and their corresponding numerical weights. It therefore evaluates their combination rather than isolating coefficient changes under fixed numerical weights; no cross-seed inference is made from this single-run contrast.

The coefficient scan compares globally fixed splits with the transition-wise rule, as discussed in Section 3 and the image experiments. The reported reward curves do not measure clipping frequency or log-ratio variance, so they do not establish which mechanism mediates the observed gains.

**Conditional gradient-scale comparison.** Fix the conditioning state and timestep pair $(t, s)$, and compare the two policies at $\theta = \theta_{\text{old}}$, where the ratio is one and clipping is inactive. Let $\boldsymbol{J} = \partial \boldsymbol{v}_\theta / \partial \theta$ be the velocity-model Jacobian. For method $m \in \{\text{CPS}, \text{Geo}\}$, the gradient of the negative, dimension-averaged policy surrogate without policy reweighting is

$$\boldsymbol{g}^{(m)}_{\text{policy}} = -\frac{A}{D} \frac{\kappa^{(m)}_{t,s}}{\alpha^{(m)}_{t,s}} \boldsymbol{J}^\top \boldsymbol{\xi} = \frac{A G^{(m)}_{t,s}}{D} \boldsymbol{J}^\top \boldsymbol{\xi}, \tag{59}$$

where the action and advantage are held fixed when differentiating, $\boldsymbol{\xi}$ is the standardized residual under the old policy, and the last equality uses $\kappa^{(m)}_{t,s} < 0$. Holding $A$, $\boldsymbol{\xi}$, $\boldsymbol{J}$, and $D$ fixed across the two policies therefore gives

$$\boldsymbol{g}^{(\text{CPS})}_{\text{policy}} = \frac{G^{\text{CPS}}_{t,s}}{G^{\star}_{t,s}} \boldsymbol{g}^{(\text{Geo})}_{\text{policy}}. \tag{60}$$

Since $G^{\text{CPS}}_{t,s} \geq G^{\star}_{t,s}$, the Flow-CPS gradient norm is no smaller than that of Geo-FSRL in this controlled comparison; the inequality is strict when the geometry factors differ and the common

gradient is nonzero. This statement concerns a fixed transition. Matching training hyperparameters alone does not hold the sampled advantages, residuals, or model Jacobians fixed, and the statement does not order the norm of a batch gradient mixing different transitions.

**Relation to the reward ordering.** For either policy, applying its matched KL weight gives

$$w_{t,s}^{\mathrm{KL}} \overline{D}_{\mathrm{KL}} = \frac{c_{\mathrm{geo}}^2}{2D} \|\boldsymbol{v}_\theta - \boldsymbol{v}_{\mathrm{ref}}\|_2^2. \tag{61}$$

Thus, retaining KL reweighting while removing policy reweighting normalizes the reference-penalty curvature in velocity space but leaves $G_{t,s}$ in the policy gradient. With the other gradient factors controlled, Flow-CPS's larger geometry factor increases the policy-gradient scale relative to this matched regularizer. This is a conditional scale explanation compatible with its slightly higher unweighted reward, rather than a measured attribution of that difference. Appendix A.4 details how full updates depend on the coefficients and optimizer. The ablation evaluates cross-timestep balancing together with changes in the overall policy-to-regularization scale.

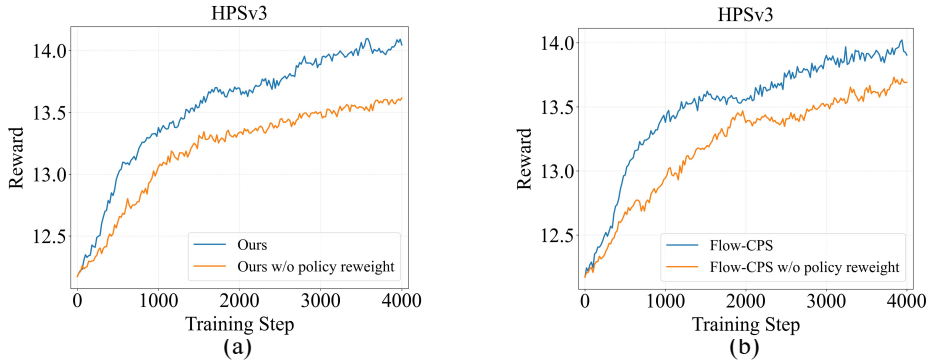


Figure 3: Validation HPSv3 with and without policy reweighting for (a) Geo-FSRL and (b) Flow-CPS. All curves use the same 256 validation prompts; horizontal axes count optimizer updates. KL reweighting remains enabled throughout. Scores are reported in Table 3.

## D Additional Qualitative Results

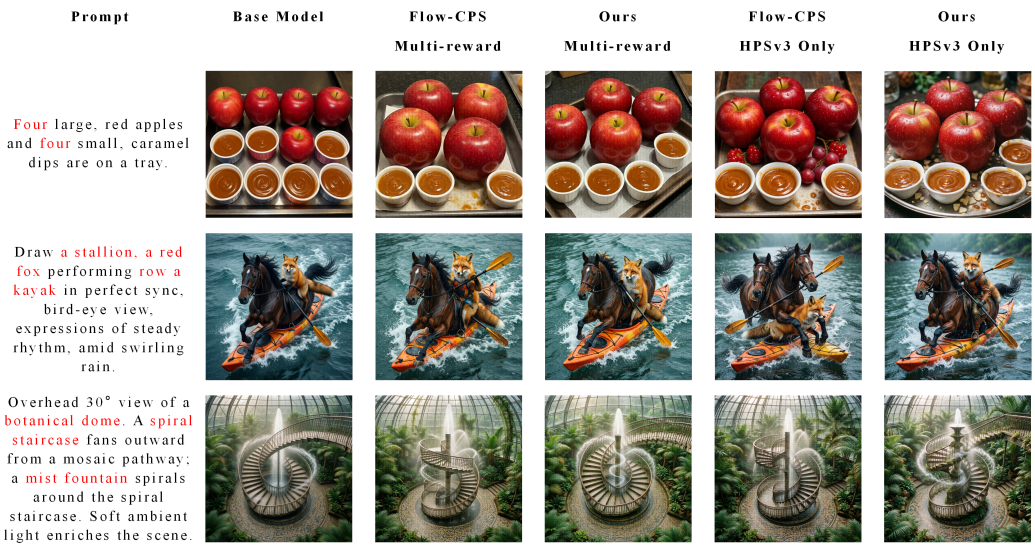


Figure 4: Qwen-Image-Lightning, Flow-CPS, and Geo-FSRL on the displayed prompts. Multi-reward training optimizes HPSv3 and CLIPScore; HPSv3-only training optimizes HPSv3 alone.

Figures 5–7 compare the base models with Geo-FSRL on additional image and video prompts.

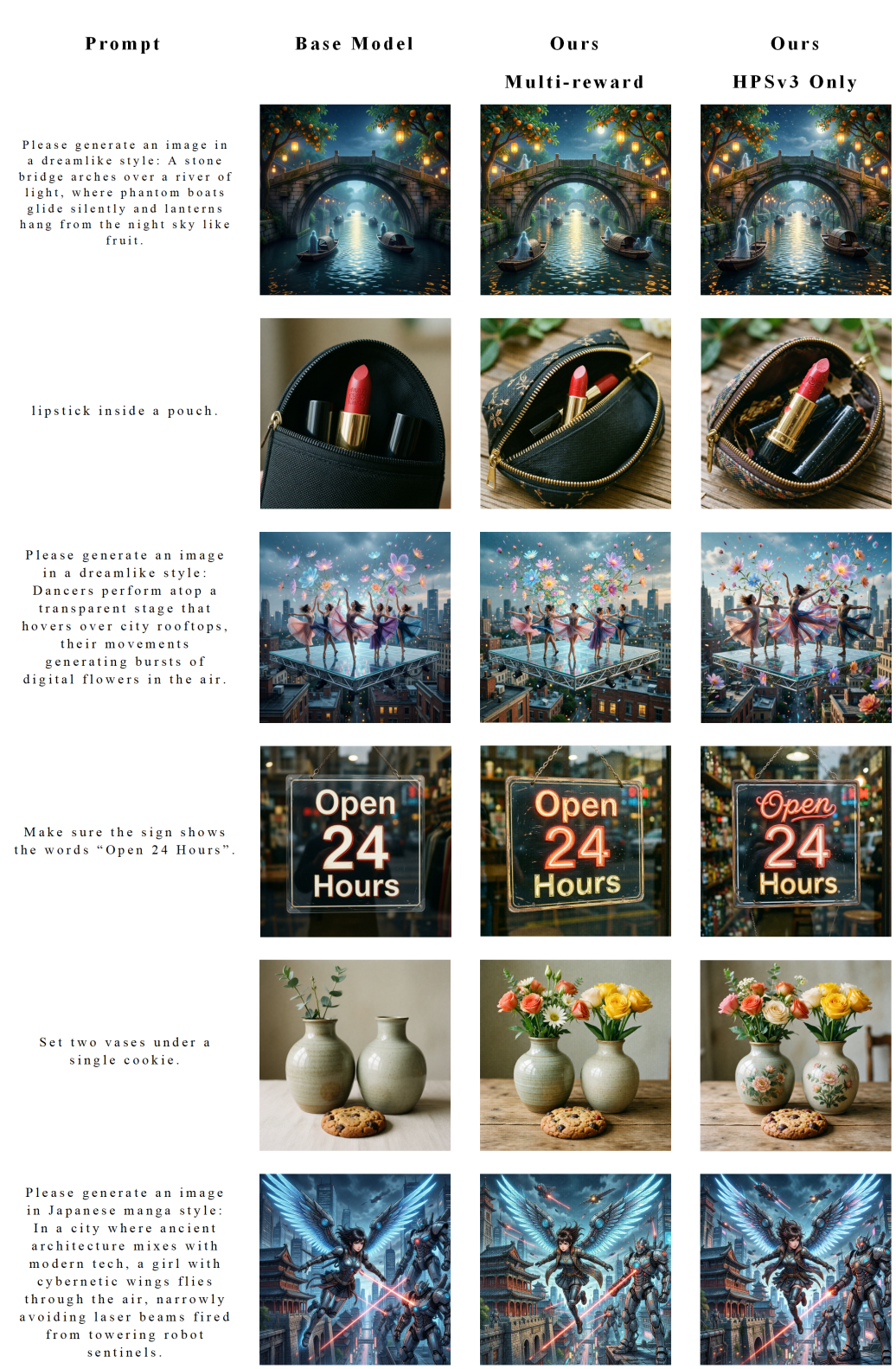


Figure 5: Additional text-to-image examples from the Qwen-Image-Lightning base model and Geo-FSRL under joint HPSv3 and CLIPScore training or HPSv3-only training. Each row uses the displayed prompt.

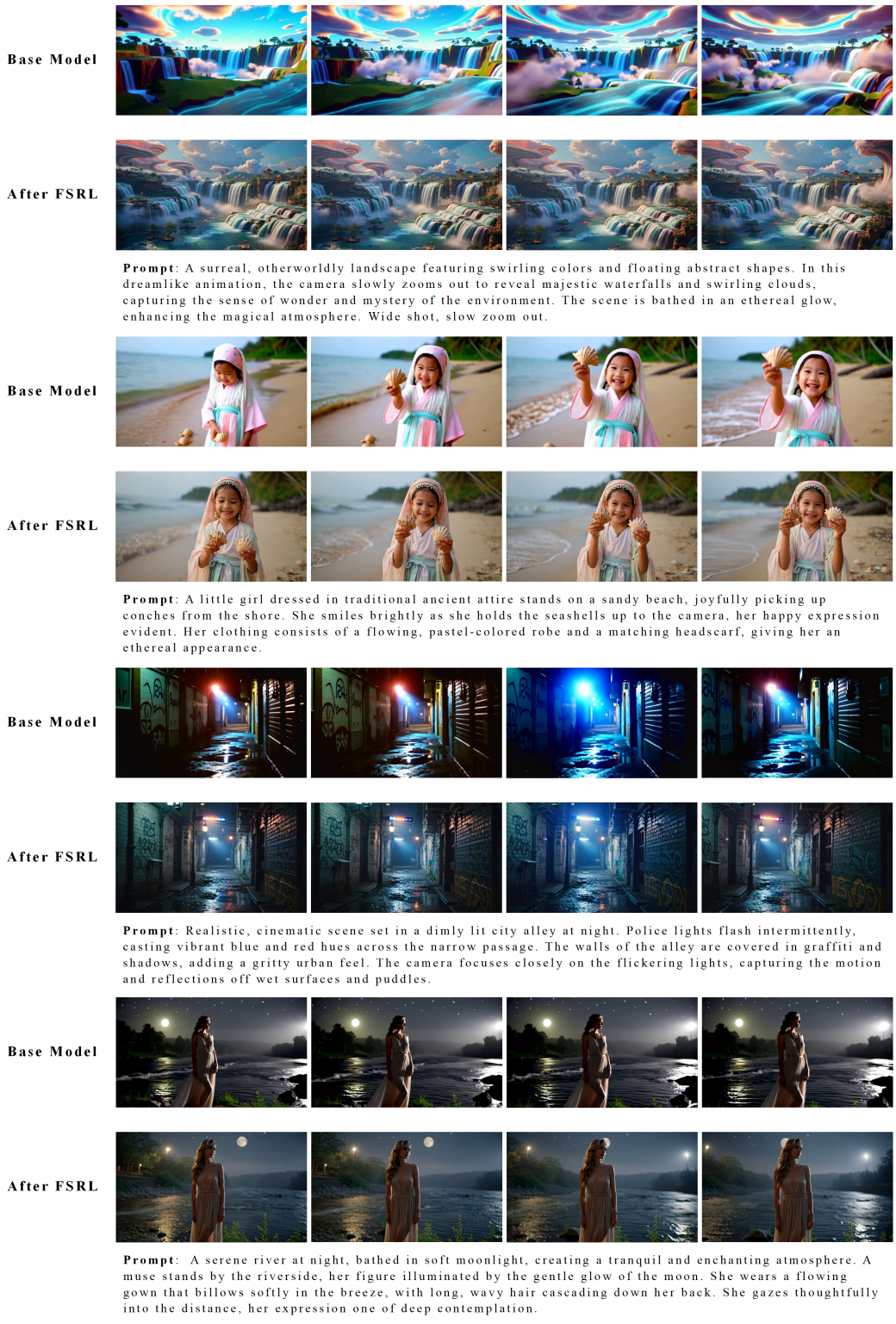


Figure 6: Selected frames from videos generated by the four-step distilled Wan2.1 base model and Geo-FSRL under joint reward optimization. For each prompt, the upper row shows the base model and the lower row ("After FSRL") shows Geo-FSRL, with four frames per video. Some prompts are shortened for display.

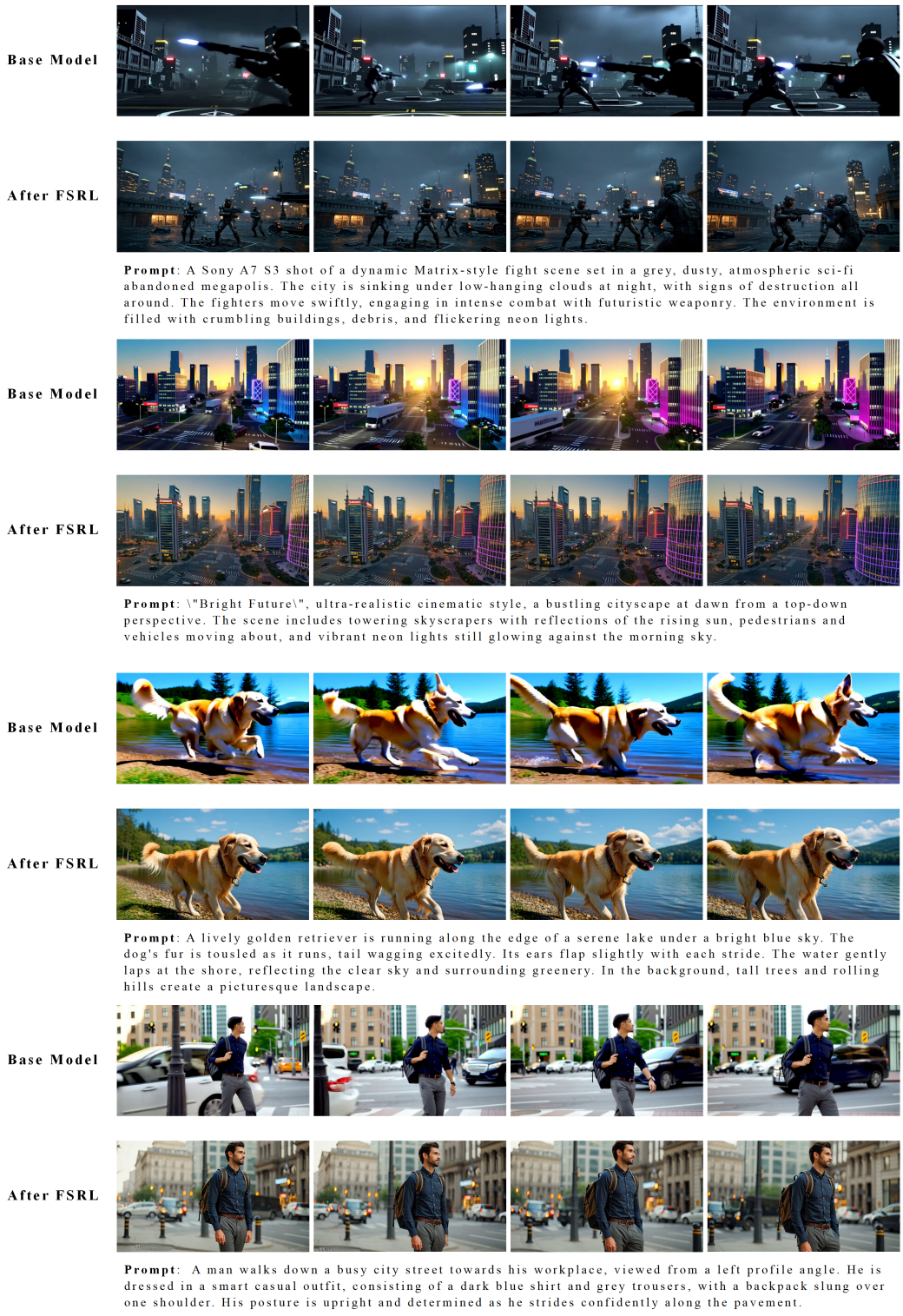


Figure 7: Additional video comparisons under joint reward optimization. Each prompt has two rows of four selected frames: the four-step distilled Wan2.1 base model above and Geo-FSRL ("After FSRL") below. Some prompts are shortened for display.