When Policy Entropy Constraint Fails: Preserving Diversity in Flow-based RLHF via Perceptual Entropy

Xiaofeng Tan[1,2,*] **Jun Liu**[2] **Bin-Bin Gao**[2] **Yuanting Fan**[2] **Xi Jiang**[3]
Chengjie Wang[2] **Hongsong Wang**[1,†] **Feng Zheng**[3,†]

[1]Southeast University [2]Tencent YouTu Lab [3]Southern University of Science and Technology
{xiaofengtan, hongsongwang}@seu.edu.cn, f.zheng@ieee.org

Abstract

RLHF is widely used to align flow-matching text-to-image models with human preferences, but often leads to severe diversity collapse after fine-tuning. In RL, diversity is often assumed to correlate with policy entropy, motivating entropy regularization. However, we show this intuition breaks in flow models: policy entropy remains constant, even while perceptual diversity collapses. We explain this mismatch both theoretically and empirically: the constant entropy arises from the fixed, pre-defined noise schedule, while the diversity collapse is driven by the mode-seeking nature of policy gradients. As a result, policy entropy fails to prevent the model from converging to a narrow high-reward region in the perceptual space. To this end, we introduce perceptual entropy that captures diversity in a perceptual space and maintains the property of standard entropy. Building upon this insight, we propose two entropy-regularized strategies, *Perceptual Entropy Constraint* (*Perc-EC*) and *Perceptual VAE Constraint* (*Perc-VC*), to preserve perceptual diversity and improve generation quality. Experiments across two base models, neural and rule-based rewards, and two perceptual spaces demonstrate consistent gains in the quality-diversity trade-off. Our project page with code is publicly available.

1 Introduction

Flow matching [37] has emerged as a powerful paradigm for text-to-image generation [78]. Since human-centric criteria such as aesthetics [15], semantic consistency [51], and safety [41] are difficult to optimize with likelihood-based pre-training alone, recent work aligns pretrained generative models through Reinforcement Learning from Human Feedback (RLHF) [18, 39, 65].

RLHF addresses this by post-training generative models with feedback from human annotators or reward models. However, diversity can decrease during fine-tuning, which limits both ***(1)*** RLHF effectiveness and ***(2)*** real-world applicability. Specifically, ***(1)*** RLHF requires diverse and informative samples for reward evaluation; reduced diversity narrows exploration and weakens feedback scope. ***(2)*** Many applications require coverage across contexts [2, 16]. As shown in Fig. 1, Flow-GRPO produces visually similar samples for both portrait and living-room prompts, while an ideal model should preserve variations in identity, pose, background, layout, lighting, and furnishing style. This motivates us to seek RLHF frameworks that *improve quality without sacrificing diversity*.

In RL research on LLMs, diversity is commonly characterized by policy entropy [11, 30, 40, 56, 90]. Under this view, recent works [12, 88] aim to balance diversity and reward $\mathcal{R}$ using policy entropy $\exp(\mathcal{H})$, and reports an empirical relationship, $\mathcal{R} = -a\exp(\mathcal{H}) + b$, with coefficients a and b. A natural solution to diversity collapse in flow-based RLHF is therefore policy entropy regularization.

*Work done during an internship at Tencent YouTu Lab. †Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Figure 1: **Diversity Collapse.** Flow-GRPO improves quality but sacrifices diversity.

However, revisiting this intuition under flow matching reveals a key inconsistency: *diversity collapses while the policy entropy surrogate remains constant*, as shown in Fig. 2. Diffusion and flow-based models [26, 57] avoid directly modeling $p(\mathbf{x}_0)$ by decomposing the process into transitions $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, whose entropy we term policy entropy. However, this quantity is fixed by the noise schedule and thus fails to reflect changes in diversity. This raises a fundamental question: *what drives diversity collapse in RL for flow models despite constant policy entropy?*

Contributions. In this paper, we investigate the diversity collapse from the perspective of entropy mechanism in flow-based RLHF, and propose a novel entropy concept with two entropy-regularization strategies to improve quality without sacrificing diversity. Main contributions are highlighted below.

First, **we investigate the apparent paradox** between constant entropy and diversity collapse in flow-based RLHF. We study two questions: ***(1)*** *why does the policy entropy remain constant after RL fine-tuning?* and ***(2)*** *why can diversity still collapse under this constant policy entropy?* ***(1)*** For the first question, we show that the noised samples $\mathbf{x}_t$ are generated by injecting noise with a *pre-defined and fixed intensity* into the model-predicted mean $\mu_\theta(\mathbf{x}_t)$, resulting in a Gaussian distribution with fixed covariance. Hence, both likelihoods and policy entropy are determined by the fixed noise scheduler. ***(2)*** For the second question, we find that this paradox stems from the mode-seeking nature of online RL, where the policy is optimized to concentrate probability distribution on high-reward region in the reward space. In this case, the multi-peak pretrained policy distribution may collapse into a single-peak perceptual distribution. This motivates an RLHF framework that encourages broad coverage of multiple high-reward regions in perceptual space, rather than convergence to a single peak, as shown in Fig. 3.

Second, motivated by the observation that policy entropy in the generative space fails to capture diversity collapse, **we introduce a novel entropy concept**, *perceptual entropy* $\mathcal{H}_{\text{perc}}$, which faithfully reflects this phenomenon while remaining consistent with prior entropy-based methods. Building on this concept, we further reveal an empirical entropy-based reward for flow models, $\mathcal{R} = -a\exp(\mathcal{H}_{\text{perc}}) + b$, which is analogous to relationships observed in LLMs [12]. Furthermore, building on the above analysis, **we propose two entropy-regularized strategies** to mitigate diversity collapse: *Perceptual Entropy Constraint* (*Perc-EC*), which incorporates perceptual entropy into the reward signal, and *Perceptual VAE Constraint* (*Perc-VC*), which aligns the perceptual distribution with the diversity-preserving VAE space to maintain diversity.

Finally, comprehensive experiments on FLUX.dev [6] and SD3.5-M [17] using neural and rule-based rewards, alongside PickScore [31], DINO [8], and CLIP [51] perceptual encoders, demonstrate that perceptual entropy improves the quality-diversity trade-off.

2 Preliminary

RLHF Objective and Policy Gradient. Given an external prompt condition $c \sim \mathcal{D}$ and a policy p_θ that generates $\mathbf{x} \sim p_\theta(\cdot \mid c)$, RLHF aims to maximize the expected reward:

$$\max_\theta \ \mathcal{J}(\theta) := \mathbb{E}_{c\sim\mathcal{D},\, \mathbf{x}\sim p_\theta(\cdot|c)}\big[\mathcal{R}(\mathbf{x}, c)\big], \tag{1}$$

Considering that the gradient of the objective is difficult to estimate directly, policy gradient methods such as REINFORCE [77] decompose it into a sum of log-likelihood gradients:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}\Big[\sum_{t=1}^{T} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)\, \mathcal{A}(\mathbf{x}, c)\Big], \tag{2}$$

where $\mathcal{A}(\mathbf{x}, c)$ denotes an advantage estimated by the final sample $\mathbf{x} \equiv \mathbf{x}_0$.

Group-Relative Advantage. GRPO [55] computes advantages via group-wise normalization. For each prompt, it samples K candidates $\{\mathbf{x}^k\}_{k=1}^K$ and normalizes rewards within the group:

$$\mathcal{A}(\mathbf{x}^i, c) = \frac{\mathcal{R}(\mathbf{x}^i, c) - \text{mean}\{\mathcal{R}(\mathbf{x}^k, c)\}}{\text{std}\{\mathcal{R}(\mathbf{x}^k, c)\}}. \tag{3}$$

Policy Density in Flow Models. For flow models, the policy in Eq. (2) is defined by the reverse-time SDE transition [18, 39]. Under the flow parameterization, each step follows a Gaussian transition

$$\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t, c) + \sigma_t \sqrt{\mathrm{d}t}\, \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \Rightarrow \quad p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = \mathcal{N}\big(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, c), \sigma_t^2 \mathrm{d}t\, \mathbf{I}\big), \tag{4}$$

where covariance $\sigma_t^2 \mathrm{d}t\, \mathbf{I}$ is fixed by noise schedule, while mean μ_θ is given by the velocity field[39]:

$$\mu_\theta(\mathbf{x}_t, t, c) = \mathbf{x}_t + \Big[v_\theta(\mathbf{x}_t, t, c) + \frac{\sigma_t^2}{2t}\big(\mathbf{x}_t + (1-t)v_\theta(\mathbf{x}_t, t, c)\big)\Big]\mathrm{d}t. \tag{5}$$

Flow-Based GRPO Objective. Flow-GRPO optimizes the policy gradient in Eq. (2) with the group-relative advantage in Eq. (3), using a clipped proximal surrogate:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{i,t}\big[\min\big(\rho_{t,i}(\theta)\mathcal{A}_i,\ \text{clip}(\rho_{t,i}(\theta), 1-\epsilon, 1+\epsilon)\mathcal{A}_i\big)\big]. \tag{6}$$

where $\rho_{t,i}(\theta) = \frac{p_\theta(\mathbf{x}_{t-1}^i | \mathbf{x}_t^i, c)}{p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}^i | \mathbf{x}_t^i, c)}$ is the probability ratio between the current and old policy.

Entropy Mechanism in LLM RL. Token-level policy entropy from $p_\theta(\mathbf{x}_t \mid \mathbf{x}_{<t})$ is widely used for entropy regularization to encourage exploration [23, 24, 30, 56, 90, 91]. In this context, Cui et al. [12] establish an empirical relationship between policy entropy and validation reward:

$$\mathcal{R} = -a \exp(\mathcal{H}) + b, \tag{7}$$

where $\mathcal{R}$ denotes the validation reward and $a, b > 0$ are coefficients.

3 Understanding the Failure of Policy Entropy Constraints

Entropy regularization is a standard mechanism for mitigating diversity collapse in RL. In flow-based RL, however, a direct entropy constraint on the generated-sample distribution is difficult to instantiate, motivating a closer examination of tractable entropy surrogates. We identify a paradox: *the policy entropy $\mathcal{H}(p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c))$ remains invariant during training, while diversity collapse still occurs*. This section analyzes this phenomenon through two questions: (1) why policy entropy remains constant during training, unlike in LLMs; and (2) why collapse still occurs despite this invariance.

3.1 From Marginal Entropy to Policy Entropy

Challenge in Estimating Marginal Entropy. In reinforcement learning, entropy regularization is typically applied to the marginal entropy $\mathcal{H}(p_\theta(\mathbf{x}_0 \mid c))$. In contrast, flow and diffusion models are trained through conditional transition distributions $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)$. As a result, evaluating $\mathcal{H}(p_\theta(\mathbf{x}_0 \mid c))$ requires access to the exact log-density of the marginal $p_\theta(\mathbf{x}_0 \mid c)$, which is widely recognized as an intractable issue in diffusion and flow model [9, 13, 14, 22, 44, 57, 73].

Joint Entropy & Policy Entropy. A natural alternative is to consider the entropy of the full trajectory $\mathbf{x}_{0:T}$ rather than the final sample $\mathbf{x}_0$. In RL for LLMs [11, 30, 40, 56, 90], the token-level policy entropy of $p_\theta(\cdot \mid \mathbf{x}_{<t})$ is widely used for regularization. Similarly, we consider using the policy entropy of $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)$ in flow models to obtain the joint entropy, as follows:

$$p_\theta(\mathbf{x}_{0:T}|c) = p(\mathbf{x}_T|c) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c), \quad (\mathbf{x}_T \perp c). \tag{8}$$

Considering that the initial distribution $\mathbf{x}_T$ follows standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbb{I})$, the joint entropy $\mathcal{H}(p_\theta(\mathbf{x}_{0:T} \mid c))$ is linearly related to the policy entropy $\mathcal{H}(p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c))$, as follow:

$$\mathcal{H}(p_\theta(\mathbf{x}_{0:T}|c)) = -\mathbb{E}_{p_\theta(\mathbf{x}_{0:T}|c)}[\log p(\mathbf{x}_T) + \sum_{t=1}^T \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)] = \mathcal{H}(p(\mathbf{x}_T)) + \sum_{t=1}^T \mathcal{H}(p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)).$$

where $\mathcal{H}(p(\mathbf{x}_T))$ is constant because $\mathbf{x}_T$ is drawn from a fixed Gaussian prior.

Thus, by analogy with token-level entropy in LLMs, the policy entropy appears to be a natural tractable surrogate for trajectory-level entropy in flow-based RL. However, we observe that it is fixed by during RL fine-tuning and therefore cannot reflect the diversity collapse observed. Below, we provide the theoretical analysis in Sec. 3.2 and verify this phenomenon empirically in Fig. 2.

3.2 Constant Policy Entropy in Flow-based RL

We now formalize the per-step policy entropy induced by the reverse Gaussian transition and show that it is determined solely by the prescribed noise schedule.

Definition 1 (Policy Entropy in Flow-based RL)**.** For a prompt $c \in \mathcal{D}$ and timestep t, the per-step conditional entropy is defined as:

$$\mathcal{H}_t(p_\theta) = -\mathbb{E}_{\{\mathbf{x}_\tau\}_{\tau=0}^{T} \sim p_\theta(\cdot|c)}[\log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)], \tag{9}$$

where the policy $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)$ is given by the Gaussian transition induced by the flow model θ, i.e., $\mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, c), \sigma_t^2\, \mathrm{d}t\, \mathbf{I})$, defined as follows:

$$\log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = -\frac{\|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t, t, c)\|^2}{2\sigma_t^2\, \mathrm{d}t} + C_t, \qquad C_t = -\tfrac{d}{2}\log\big(2\pi\sigma_t^2\, \mathrm{d}t\big), \tag{10}$$

where d denotes the dimensionality of $\mathbf{x}_t$, and C_t is constant term that is independent of θ.

Definition 1 provides the closed-form density of the policy probability and its entropy. To further analyze the quantitative property of policy entropy, we provide the following property.

Property 1 (Constant Policy Entropy in Flow-based RL)**.** For the reverse-time transition in Eq. (4), the per-step conditional entropy at timestep t satisfies (see in App. B.1):

$$\mathcal{H}_t(p_\theta) = \frac{d}{2} - C_t. \tag{11}$$

Property 1 shows that the policy entropy is invariant to the model parameters θ and depends only on the noise schedule and dimensionality. Intuitively, variations in θ only affect the Gaussian mean $\mu_\theta(\mathbf{x}_t, t, c)$ while leaving the covariance $\sigma_t^2\, \mathrm{d}t\, \mathbf{I}$ unchanged. Thus, the policy entropy remains constant during model θ fine-tuning, since the covariance is fixed by the noise schedule.

Beyond the theoretical analysis, we empirically verify the constant policy entropy property in Fig. 2. As shown in Fig. 2(a), during training, the entropy and variance in the VAE space remain nearly invariant. ***However, diversity in the reward-perceived space still decreases over training.*** Consistent trends are also observed in the pixel and reward spaces, as shown in Fig. 2(b). Next, we analyze the underlying reason for this apparent contradiction.

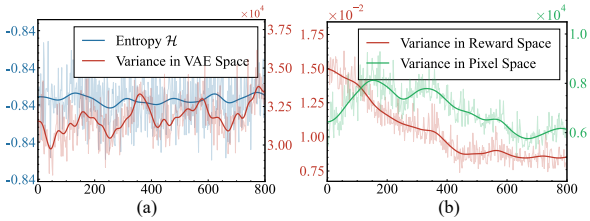

Figure 2: **Training-time Metrics.** (a) Entropy vs. variance in VAE space. (b) Variance in pixel and reward space.

3.3 Diversity Collapse behind Constant Policy Entropy

The previous section explains why entropy stays constant. We next examine why diversity still collapses. Starting from the policy-gradient objective in Eq. (2), the following corollary characterizes the local policy-gradient direction underlying GRPO, with proof deferred to App. B.3.

Corollary 1 (Mode-Seeking Optimization under On-Policy Policy Gradient)**.** *For policy gradient methods [77] in Eq. (2), with $p_{\theta_{\text{old}}}$ and $\mathcal{A}$ fixed within an update, the gradient satisfies:*

$$\nabla_\theta \mathcal{J}(\theta) = \nabla_\theta\Big(- D_{\text{KL}}\big(p_\theta(\mathbf{x}_{0:T}|c) \,\|\, p_{\mathcal{A}}(\mathbf{x}_{0:T}|c)\big) + D_{\text{KL}}\big(p_\theta(\mathbf{x}_{0:T}|c) \,\|\, p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c)\big)\Big),$$

where $p_\theta(\mathbf{x}_{0:T}|c)$ is the joint distribution, and $p_{\mathcal{A}}(\mathbf{x}_{0:T}|c)$ is the target distribution induced by advantage on the same trajectory space $p_{\mathcal{A}}(\mathbf{x}_{0:T}|c) = Z_{\mathcal{A}}(c)^{-1} p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c)\exp(\mathcal{A}(\mathbf{x}_0, c))$ with $Z_{\mathcal{A}}(c) = \int p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c)\exp(\mathcal{A}(\mathbf{x}_0, c))\, \mathrm{d}\mathbf{x}_{0:T}$.

Corollary 1 shows that the unclipped update direction locally minimizes the KL divergence to the advantage-induced target $p_{\mathcal{A}}$ while the KL term w.r.t. $p_{\theta_{\text{old}}}$ acts as a regularizer, with GRPO applying trust-region clipping to this update. Since both KL terms place p_θ as the first argument, they correspond to the reverse KL, which is mode-seeking [29] and drives the policy toward high-reward regions in the perceptual space, while being regularized toward the old policy in the generation space.

Conceptual Illustrations. Intuitively, the target distribution $p_{\mathcal{A}}$ is sampled from the old policy θ_{old}, and thus unexplored perceptual modes are excluded from optimization. As a result, the policy model

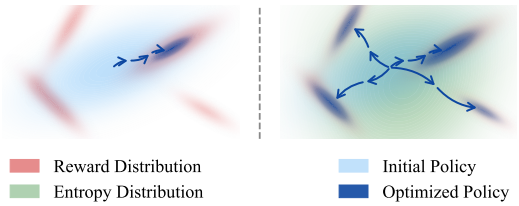

Figure 3: **Optimization Behavior.** GRPO shows mode-seeking behavior and converges to a single-peak distribution, while an ideal model should encourage coverage of multiple regions.

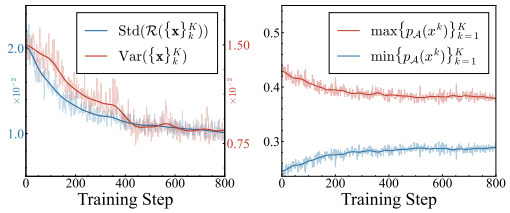

Figure 4: **Training Dynamics Metrics.** (a) standard deviation of reward and variance of samples; (b) maximum and minimum advantage probabilities $p_{\mathcal{A}}$ during training.

is encouraged toward a single high-reward mode with limited coverage, leading to diversity collapse. Repeated on-policy updates therefore converge probability mass on a narrow high-reward region in the perceptual space, as shown in Fig. 3. In contrast, an ideal model augmented with a perceptual signal should promote coverage over multiple high-reward regions. Empirically, Fig. 4 shows that the dispersion of $p_{\mathcal{A}}$ decreases during training, with both reward variance and sample diversity declining, consistent with the contraction predicted by Corollary 1.

Summary. The policy entropy in the generation space remains invariant, as its covariance is fixed by the noise schedule; however, mode-seeking optimization drives the model to concentrate on a single high-reward mode with low perceptual diversity.

4 Method

In this section, we address the issue of diversity collapse. Intuitively, diversity collapse can be seen as the feature variance reduction in human perception. Our analysis above reveals a key factor: ***the standard policy entropy in flow models is unable to capture this collapse***, undermining the effectiveness of existing entropy-based regularization for flow matching. This motivates us to define a new entropy concept, *perceptual entropy*, which captures diversity collapse in perceptual space while remaining consistent with existing entropy-based regularization methods.

> **Insight (Representation Perspective).** Standard VAEs primarily extract low-level features, making VAE-space entropy $\mathcal{H}$ insensitive to semantic diversity. For example, slightly altering eye size in a portrait shifts VAE latents while leaving high-level semantics nearly unchanged. In contrast, perceptual entropy $\mathcal{H}_{\text{perc}}$ in a high-level representation space directly captures semantic variations and reliably detects diversity collapse.

4.1 Perceptual Entropy

To capture sample diversity via policy entropy, we introduce the concept of *perceptual entropy*. It preserves the structure of standard policy entropy but is defined in the perceptual space rather than the generative latent space, as follows.

Definition 2 (Perceptual Entropy)**.** Let $\phi = \mathcal{D}_{\text{vae}} \circ \mathcal{E}_p$ be a mapping where $\mathcal{E}_p$ is a perceptual encoder and $\mathcal{D}_{\text{vae}}$ is a VAE decoder. Given k rollout noised samples from the exploration process Eq. (4), the perceptual probability is defined as:

$$p_\theta^{\text{perc}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c) := \mathcal{N}\Big(\phi(\mathbf{x}_{t-1}^k);\, \phi(\mathbf{m}^k),\, \sigma_t^2\, \mathrm{d}t\, \mathbf{I}\Big)\,. \tag{12}$$

where $\mathbf{m}^k$ denotes the old-policy mean $\mu_{\theta_{\text{old}}}(\mathbf{x}_t^k, t, c)$, and the perceptual entropy is then:

$$\mathcal{H}_{\text{perc}}(p_\theta) = \mathbb{E}_{t,k}\Big[-\log p_\theta^{\text{perc}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c)\Big]\,. \tag{13}$$

Definition 2 introduces an entropy notion that quantifies diversity collapse in the perceptual space. Specifically, noised samples $\mathbf{x}_t$ and the old-policy mean $\mathbf{m}^k$ in the flow model are first decoded into the pixel space by a VAE decoder $\mathcal{D}_{\text{vae}}$, and then mapped into the perceptual space by a perceptual encoder $\mathcal{E}_p$. In this space, perceptual likelihoods $p_\theta^{\text{perc}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c)$ are computed under a local Gaussian centered at $\phi(\mathbf{m}^k)$, and used to define the perceptual entropy in Eq. (13).

Perceptual Encoder. Here, we discuss the choice of perceptual encoder. For neural network-based rewards, we use the PickScore encoder to avoid additional computational cost. As discussed in Sec. 5.3 and Tab. 4, this additional perceptual encoder is both effective and efficient, yielding consistent gains with minimal overhead.

Surrogate Scoring without Jacobians. Definition 2 formulates p_θ^{perc} directly as a feature-space scoring distribution without the spatial-transformation Jacobian, analogous to score distillation methods that optimize 2D priors directly in feature space [36, 48, 71, 74]. Following on-policy REINFORCE formulations [1, 27, 42], perceptual scores are evaluated on fixed old-policy rollouts as shaped rewards, while the GRPO likelihood ratio performs the policy update.

To connect perceptual entropy to sample diversity, we analyze its relationship to the conditional feature-space variance under a local mean-preserving approximation.

Remark 1 (*Connection to Conditional Feature-Space Variance*). Let $\mathbf{z}_{t-1} = \phi(\mathbf{x}_{t-1})$. Assume that, within a local reverse step, the perceptual encoder is approximately mean-preserving, i.e., $\mathbb{E}[\mathbf{z}_{t-1} \mid \mathbf{x}_t] \approx \phi(\mathbf{m})$. Then perceptual entropy tracks the expected conditional feature-space variance:

$$\mathcal{H}_{\text{perc}}(p_\theta) \approx \mathbb{E}_{t,\mathbf{x}_t}\left[\frac{\text{Var}(\mathbf{z}_{t-1} \mid \mathbf{x}_t)}{2\sigma_t^2 \mathrm{d}t}\right] - C_t. \tag{14}$$

Remark 1 shows that perceptual entropy captures local perceptual dispersion, paralleling the latent-space variance relation in App. B.2 and aligning with entropy-regularization analyses in LLMs [12, 30].

Empirical Entropy Mechanism. To better understand this mechanism, we empirically study the relationship between perceptual entropy and reward in flow models, in analogy to the empirical finding in LLMs [12]. We therefore propose a similar empirical entropy mechanism for flow models:

$$\mathcal{R} = -a \exp(\mathcal{H}_{\text{perc}}) + b. \tag{15}$$

where a and b are empirical coefficients.

As shown in Fig. 5, we report the relationship between reward $\mathcal{R}$ and perceptual entropy $\mathcal{H}_{\text{perc}}$. These results along with the theoretical analyses demonstrate that perceptual entropy provides ***an effective identifier of diversity***, ***aligning with existing research*** for LLMs [12].

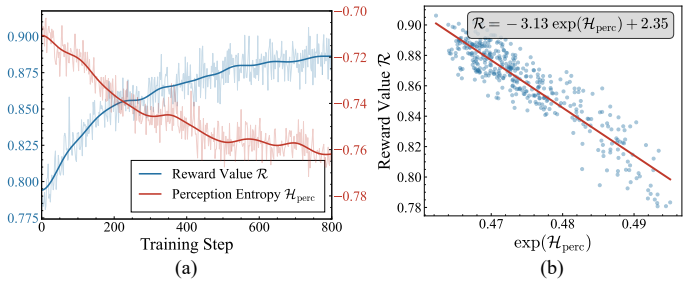

Figure 5: **Relationship between Reward and Perceptual Entropy.**(a) Reward $\mathcal{R}$ vs. perceptual entropy $\mathcal{H}_{\text{perc}}$. (b) Reward $\mathcal{R}$ vs. perceptual entropy $\mathcal{H}_{\text{perc}}$.

This naturally motivates regularizing perceptual entropy to mitigate diversity collapse. Here, we introduce two approaches based on perceptual entropy: (1) *Perceptual Entropy Constraint* (*Perc-EC*), which treats perceptual entropy as a principled entropy identifier and incorporates it into a standard entropy-regularization framework; and (2) *Perceptual VAE Constraint* (*Perc-VC*), which leverages perceptual entropy to encourage consistency between the original and collapsed perceptual spaces.

4.2 Perceptual-Regularized Strategies for GRPO

Objective. Standard RLHF inherently drives the policy toward a narrow subset of high-reward samples, thereby inducing diversity collapse. To mitigate this issue, we augment the RLHF objective in Eq. (1) with a perceptual regularization term $\Omega(p_\theta^{\text{perc}})$, as follows:

$$\max_\theta \ \mathcal{J}(\theta) = \mathbb{E}_{\mathbf{x}\sim p_\theta}\left[\mathcal{R}(\mathbf{x}, c)\right] + \lambda\, \Omega(p_\theta^{\text{perc}}). \tag{16}$$

Here, $\Omega(p_\theta^{\text{perc}})$ is a perceptual regularizer that encourages exploration of a broader perceptual space, while $\mathbb{E}_{\mathbf{x}\sim p_\theta}\left[\mathcal{R}(\mathbf{x}, c)\right]$ promotes high-preference samples through reward maximization.

Perceptual Entropy Constraint (*Perc-EC*). Since perceptual entropy serves as a reliable surrogate for conditional variance as shown in Remark 1, a natural choice is to maximize it directly, as follows:

$$\max_\theta \ \mathcal{J}_{\textit{Perc-EC}}(\theta) = \mathbb{E}_{\mathbf{x}\sim p_\theta}\left[\mathcal{R}(\mathbf{x}, c)\right] + \lambda\, \mathcal{H}_{\text{perc}}(p_\theta). \tag{17}$$

Here, we set $\Omega = \mathcal{H}_{\text{perc}}(p_\theta)$ to encourage the policy to explore a broader perceptual space.

Perceptual VAE Constraint (*Perc-VC*). While *Perc-EC* directly promotes variance, an alternative is to anchor the policy distribution to the diversity-preserving generation-space prior by setting $\Omega = -D_{\text{KL}}(p_\theta^{\text{perc}} \| p_\theta^{\text{vae}})$, where p_θ^{vae} is the policy p_θ in Eq. (4), as follows:

$$\max_\theta \; \mathcal{J}_{\textit{Perc-VC}}(\theta) = \mathbb{E}_{\mathbf{x}\sim p_\theta}\big[\mathcal{R}(\mathbf{x}, c)\big] - \lambda\, D_{\text{KL}}(p_\theta^{\text{perc}} \,\|\, p_\theta^{\text{vae}})\,. \tag{18}$$

Perceptual Reward Shaping. Although theoretically sound, optimizing *Perc-EC* and *Perc-VC* as independent auxiliary losses introduces empirical challenges. Direct entropy maximization injects unstructured stochasticity [64], leading to competing gradients against the primary reward and instability training, as shown in the red line in Fig. 6(a). To address this, we incorporate perceptual constraints directly into the reward signal, a strategy we term *perceptual reward shaping*. The shaped rewards are computed by aggregating instantaneous perceptual quantities across reverse timesteps:

$$\mathcal{J}_{\textit{Perc-EC}}(\theta) = \mathbb{E}_{c,k}\Big[\underbrace{\mathcal{R}(\mathbf{x}_0^k, c) - \lambda\, \mathbb{E}_t\Big[\log p_\theta^{\text{perc}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c)\Big]}_{\text{Shaped Reward } \tilde{\mathcal{R}}^k_{\textit{Perc-EC}}}\Big]. \tag{19}$$

This incorporates perceptual constraints directly into the reward signal, ensuring that perceptual and task rewards are jointly normalized within the same GRPO rollout group, as shown by the blue line in Fig. 6(a). Similarly, the shaped reward for the *Perc-VC* constraint is given by:

$$\tilde{\mathcal{R}}^k_{\textit{Perc-VC}} = \mathcal{R}(\mathbf{x}_0^k, c) - \lambda\, \mathbb{E}_t\Big[\log p_\theta^{\text{perc}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c) - \log p_\theta^{\text{vae}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c)\Big]\,. \tag{20}$$

To optimize the shaped rewards, we adopt the standard GRPO clipped surrogate, as follows:

$$\mathcal{J}_{m\text{-GRPO}}(\theta) = \mathbb{E}_{c,\,\{\mathbf{x}_t^k\}_{t,k}}\Big[\min\big(\rho_{t,k}(\theta)\,\hat{\mathcal{A}}_k^m,\; \text{clip}(\rho_{t,k}(\theta), 1-\epsilon, 1+\epsilon)\,\hat{\mathcal{A}}_k^m\big)\Big]. \tag{21}$$

where $\hat{\mathcal{A}}_k^m$ is the advantage in Eq.(3), and $m \in \{\textit{Perc-EC}, \textit{Perc-VC}\}$ denotes the constraint type.

As illustrated in Fig.6, our method effectively mitigates these optimization issues. Compared to Native KL, which introduces competing gradients, perceptual reward shaping aligns perceptual constraints with the primary objective and stabilizes training dynamics, as shown by the blue line in Fig.6(a). Moreover, incorporating the *Perc-VC* constraint regularizes entropy maximization, leading to higher and more stable entropy and improved output diversity, as shown in Fig. 6(b).

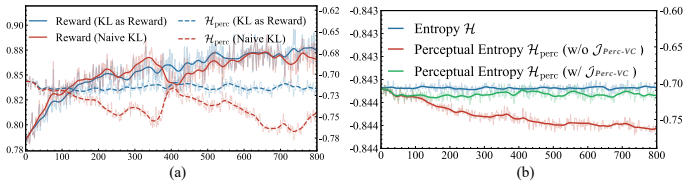

Figure 6: **Optimization Comparison.** (a) Native KL (red) versus KL-as-reward (blue), showing reward and entropy dynamics; (b) $\mathcal{H}$ (blue) and $\mathcal{H}_{\text{perc}}$ with (green) and without (red) $\mathcal{J}_{\textit{Perc-VC}}$.

5 Experiments

5.1 Experimental Setup

Implementation. We build upon Flow-GRPO [39] and evaluate two base models, FLUX.dev [6] and SD3.5-M [17]. For FLUX.dev, we use PickScore [31] as the neural reward and compute perceptual entropy with PickScore and DINO [8] encoders. For SD3.5-M, we follow Flow-GRPO [39], which adopts rule-based rewards. FLUX.dev is trained on Pick-a-Pic [31] with 37,523 prompts. For *Perc-EC* and *Perc-VC*, we tune $\lambda \in \{0.03, 0.05, 0.10\}$, with larger values to encourage diversity and smaller ones for quality. We also apply *Perc-EC* and *Perc-VC* to DDPO [5] on FLUX.dev with PickScore, and train Wan2.1-T2V-1.3B [62] for text-to-video generation with Flow-GRPO and VideoPickScore in UniRL [67]. Image generation models are trained on 8 GPUs, and video generation models are trained on 96 GPUs.

Metrics. Following MixGRPO [33], we evaluate generation quality on the HPD test set using ImageReward [82], PickScore [31], Aesthetic Predictor v2.5 [15], CLIP [51], HPSv2.1 [79], and Unified Reward [72]. For diversity, we sample 100 images per prompt across 85 PartiPrompts [86], totaling 8,500 images per method, and report feature-space L_2 diversity in DINO and CLIP spaces

Table 1: **Entropy Regularization Comparison.** Gray-shaded rows denote methods using perceptual entropy $\mathcal{H}_{\text{perc}}$. Abbreviations: P.S.=PickScore, I.R.=ImageReward, U.R.=Unified Reward, and V.S.=Vendi Score.

Method	Perceptual Space	Quality↑							Diversity↑							Overall
		P.S.	Aes.v2.5	I.R.	CLIP	HPSv2.1	U.R.	Norm.Avg	DINO	CLIP	V.S.$_{\text{DINO}}$	V.S.$_{\text{CLIP}}$	V.S.$_{\text{IR}}$	V.S.$_{\text{PS}}$	Norm.Avg	
Base Model & Baseline without Entropy Regularization																
Flux Dev	–	0.226	5.837	1.089	0.388	0.312	3.540	0.000	0.70	0.49	5.14	2.61	1.26	1.15	0.389	0.195
Flow-GRPO (*Baseline*)	–	0.243	6.305	1.375	0.396	0.337	3.738	0.685	0.54	0.35	2.82	1.69	1.13	1.09	0.010	0.347
Existing Methods with Native Entropy $\mathcal{H}$ or Perceptual Entropy $\mathcal{H}_{\text{perc}}$																
Flow-GRPO w/ KL	–	0.233	5.998	1.289	0.396	0.323	3.582	0.319	0.61	0.41	3.64	2.02	1.18	1.12	0.165	0.242
Clip-Higher	–	0.240	6.273	1.367	0.392	0.336	3.741	0.617	0.55	0.38	3.01	1.90	1.17	1.10	0.066	0.342
Entropy Reg.	VAE	0.240	5.965	1.362	0.398	0.332	3.615	0.473	0.53	0.37	2.74	1.74	1.15	1.09	0.018	0.245
Entropy Reg.	PickScore	0.239	6.079	1.365	0.396	0.332	3.604	0.478	0.69	0.51	4.52	2.58	1.27	1.16	0.376	0.427
Clip-Cov	VAE	0.241	6.128	1.426	0.399	0.335	3.672	0.610	0.59	0.41	3.42	1.98	1.15	1.11	0.121	0.366
Clip-Cov	PickScore	0.242	6.153	1.448	0.402	0.337	3.698	0.680	0.60	0.42	3.49	2.11	1.19	1.12	0.163	0.422
KL-Cov	VAE	0.243	6.143	1.431	0.397	0.345	3.720	0.694	0.59	0.39	3.47	1.86	1.15	1.11	0.112	0.403
KL-Cov	PickScore	0.243	6.296	1.474	0.398	0.346	3.746	0.783	0.59	0.41	3.42	1.98	1.15	1.11	0.121	0.452
The Proposed Methods with Perceptual Entropy $\mathcal{H}_{\text{perc}}$ from DINO Space																
Perc-VC (Ours, $\lambda = 0.03$)	DINO	0.244	6.362	**1.497**	0.405	0.344	3.774	0.876	0.64	0.43	4.06	2.15	1.21	1.14	0.241	0.559
Perc-VC (Ours, $\lambda = 0.10$)	DINO	0.237	6.176	1.328	0.401	0.332	3.665	0.546	**0.78**	0.61	**6.88**	3.94	1.35	1.21	0.680	0.613
Perc-EC (Ours, $\lambda = 0.05$)	DINO	0.238	6.405	1.428	0.406	0.336	3.717	0.746	0.70	0.52	4.83	2.86	1.30	1.16	0.420	0.583
Perc-EC (Ours, $\lambda = 0.10$)	DINO	0.233	6.195	1.327	0.396	0.332	3.640	0.468	**0.78**	0.67	6.72	5.18	1.41	1.27	0.802	0.635
The Proposed Methods with Perceptual Entropy $\mathcal{H}_{\text{perc}}$ from PickScore Space																
Perc-VC (Ours, $\lambda = 0.03$)	PickScore	**0.246**	6.388	1.489	0.408	**0.348**	**3.816**	**0.962**	0.61	0.44	3.61	2.17	1.22	1.13	0.204	0.583
Perc-VC (Ours, $\lambda = 0.10$)	PickScore	0.240	6.211	1.335	0.397	0.335	3.683	0.581	0.73	**0.80**	5.44	**7.33**	1.42	1.31	0.862	0.721
Perc-EC (Ours, $\lambda = 0.05$)	PickScore	0.240	**6.412**	1.446	**0.412**	0.338	3.748	0.842	0.66	0.52	4.10	2.71	1.26	1.15	0.335	0.589
Perc-EC (Ours, $\lambda = 0.10$)	PickScore	0.235	6.242	1.340	0.401	0.334	3.674	0.568	**0.78**	0.74	6.41	6.73	**1.51**	**1.32**	**0.941**	**0.755**

Table 2: **SoTA Comparison.** Type denotes post-training: N=None, U=Unknown, D=DPO, and R=reinforcement fine-tuning.

Method	Type	Quality↑							Diversity↑							Overall
		PickScore	Aes.v2.5	ImageReward	CLIP	HPSv2.1	Unified Reward	Norm.Avg	DINO	CLIP	V.S.$_{\text{DINO}}$	V.S.$_{\text{CLIP}}$	V.S.$_{\text{IR}}$	V.S.$_{\text{PS}}$	Norm.Avg	
SDXL [47]	N	0.226	5.836	0.980	0.411	0.294	3.410	0.205	**0.83**	0.54	**8.50**	3.06	1.50	1.14	0.651	0.428
Hunyuan-DiT [35]	U	0.224	5.659	1.080	0.406	0.303	3.428	0.190	0.74	0.51	6.22	2.92	1.32	1.13	0.438	0.314
SD-3.5-M [17]	D	0.227	5.711	1.113	0.404	0.301	3.553	0.256	0.76	0.51	6.78	2.85	1.33	1.13	0.467	0.361
Kolors [32]	N	0.225	5.994	0.977	0.385	0.314	3.362	0.143	0.78	0.51	7.59	2.83	1.40	1.16	0.550	0.347
SD-3.5-L [17]	D	0.229	5.897	1.122	0.408	0.302	3.619	0.359	0.71	0.44	5.53	2.25	1.22	1.11	0.299	0.329
HiDream-Dev [7]	U	0.229	6.049	1.329	0.398	0.325	3.781	0.527	0.57	0.39	3.58	2.08	1.19	1.10	0.124	0.326
Qwen-Image [78]	D&R	0.232	6.039	1.409	**0.417**	0.324	**3.921**	0.711	0.52	0.35	2.77	1.78	1.12	1.07	0.000	0.356
Flux Dev *(Base Model)*	U	0.226	5.837	1.089	0.388	0.312	3.540	0.215	0.70	0.49	5.14	2.61	1.26	1.15	0.356	0.285
Perc-VC (Ours, $\lambda = 0.03$)	R	**0.246**	6.388	**1.489**	0.408	**0.348**	3.816	**0.917**	0.61	0.44	3.61	2.17	1.22	1.13	0.201	0.559
Perc-VC (Ours, $\lambda = 0.10$)	R	0.240	6.211	1.335	0.397	0.335	3.683	0.645	0.73	**0.80**	5.44	**7.33**	1.42	1.31	0.812	0.729
Perc-EC (Ours, $\lambda = 0.05$)	R	0.240	**6.412**	1.446	0.412	0.338	3.748	0.832	0.66	0.52	4.10	2.71	1.26	1.15	0.318	0.575
Perc-EC (Ours, $\lambda = 0.10$)	R	0.235	6.242	1.340	0.401	0.334	3.674	0.630	0.78	0.74	6.41	6.73	**1.51**	**1.32**	**0.872**	**0.751**

alongside Vendi Scores [19] in CLIP, DINO, ImageReward, and PickScore spaces. We also report an overall score by averaging min-max normalized quality and diversity metrics within each table.

Compared Methods. We compare standard entropy regularization [54], Clip-Higher [87], and covariance-aware Clip-Cov/KL-Cov [12]. Details are provided in App. C. We test both their original entropy terms and our perceptual entropy $\mathcal{H}_{\text{perc}}$, then compare directly with *Perc-VC* and *Perc-EC*. Hyperparameters follow Flow-GRPO unless stated otherwise: entropy weight 0.05, Clip-Cov/KL-Cov rate 0.25, Clip-Higher threshold 0.1, and KL weight 0.001 [39].

5.2 Main Results

Comparison with Entropy-Based Regularization. As shown in Tab. 1, native per-step entropy $\mathcal{H}$ provides limited gains under flow matching, whereas perceptual entropy $\mathcal{H}_{\text{perc}}$ is substantially more effective. Standard entropy regularization achieves an overall score of 0.245, comparable to Flow-GRPO with KL at 0.242. Replacing $\mathcal{H}$ with $\mathcal{H}_{\text{perc}}$ raises the score to 0.427. *Perc-EC* ($\lambda = 0.10$, PickScore) achieves the best overall score of 0.755 and diversity average of 0.941.

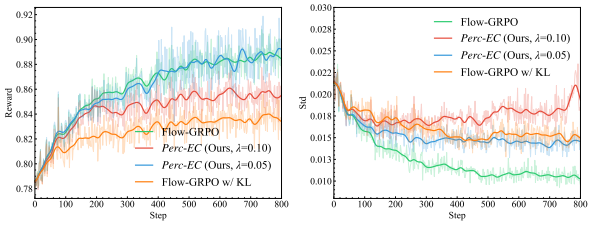

Figure 7: **Training Dynamics of Flow-GRPO, Flow-GRPO with KL, and *Perc-EC* with Perceptual Entropy.** (a) Mean reward. (b) Reward standard deviation.

Comparison with SoTA Models. As shown in Tab. 2, existing text-to-image models exhibit a sharp quality–diversity trade-off. SDXL achieves the highest baseline diversity of 0.651 but a lower quality

Table 3: **Perceptual Entropy Ablation.** Our perceptual entropy $\mathcal{H}_{\text{perc}}$ for *Perc-EC* is replaced with native entropy $\mathcal{H}$.

Method	Entropy	Aes.v2.5↑	ImageReward↑	CLIP↑	DINO↑
Perc-EC	$\mathcal{H}$	6.372	1.439	0.405	0.61
Perc-EC	$\mathcal{H}_{\text{perc}}$	**6.412**	**1.446**	**0.412**	**0.66**

Table 4: **Computational Cost.** Iteration time is normalized by Flow-GRPO. Cost is limited.

Method	Perceptual Encoder	Iteration Time	Peak Memory
Flow-GRPO	–	1.00×	78.7G
Perc-EC (Ours)	PickScore Encoder	1.01×	78.7G
Perc-EC (Ours)	DINO Encoder	1.03×	79.2G

of 0.205, whereas Qwen-Image attains higher quality of 0.711 with substantially reduced diversity of 0.000. In contrast, *Perc-EC* ($\lambda = 0.10$, PickScore) reaches the best overall score of 0.751 with a diversity average of 0.872.

Trade-Off between Diversity and Quality. Fig. 7 shows that Flow-GRPO improves quality but reduces diversity. KL alleviates this drop at a quality cost. *Perc-EC* gives a better balance: $\lambda = 0.10$ preserves diversity, while $\lambda = 0.05$ keeps quality close to Flow-GRPO.

Visualization & User Study. The teaser in Fig. 1 and extended visualizations in Fig. A1 demonstrate that *Perc-EC* preserves visual diversity across prompts and domains while maintaining high visual fidelity. User study details are provided in App. A.1.

5.3 Ablation Study

Effectiveness of Perceptual Entropy. Tab. 1 reports KL-Cov at 0.403 overall in VAE space and 0.452 in PickScore space. Tab. 3 further isolates this effect in *Perc-EC*, where replacing the native per-step entropy with $\mathcal{H}_{\text{perc}}$ improves Aes.v2.5 from 6.372 to 6.412.

Perceptual Reward Shaping vs. Native Regularization. Fig. 6 shows that perceptual reward shaping is more stable than directly optimizing the native KL or entropy regularization. This stabilizer improves quality while preserving the diversity benefit from perceptual entropy.

Perceptual Encoder. Our perceptual entropy can be computed not only with the reward model encoder, but also with a frozen DINO or CLIP encoder. As shown in Table 1, the performance gains generalize across different perceptual encoders. As shown in Table 4, the additional iteration-time overhead is marginal and peak memory usage is nearly unchanged, since the perceptual encoder is much smaller than the flow model and is evaluated asynchronously from generation and optimization.

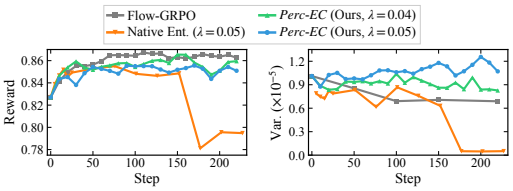

Figure 8: **Text-to-Video Generalization.** Training dynamics on Wan2.1. **Left:** Reward $\mathcal{R}$. **Right:** Within-group perceptual variance.

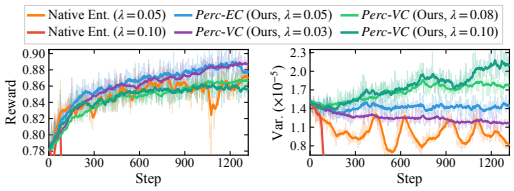

Figure 9: **Generalization to DDPO.** Training dynamics under DDPO. **Left:** Reward $\mathcal{R}$. **Right:** Within-group perceptual variance.

Text-to-Video Generalization. As shown in Fig. 8, both Flow-GRPO and native entropy regularization erode within-group perceptual variance on Wan2.1 as reward rises, whereas *Perc-EC* preserves perceptual diversity alongside high reward.

Generalization to DDPO. As shown in Fig. 9, native entropy regularization fails to prevent diversity collapse under DDPO: variance still declines at $\lambda = 0.05$, while $\lambda = 0.10$ quickly destabilizes training. In contrast, *Perc-EC* maintains perceptual variance alongside high reward, and *Perc-VC* enables a controllable quality–diversity trade-off that progressively expands perceptual variance.

Per-Prompt Diversity. Fig. 10 reports the per-prompt Vendi Scores across the top-50 PartiPrompts with 100 outputs per prompt in CLIP, DINO, and DINOv2 feature spaces.

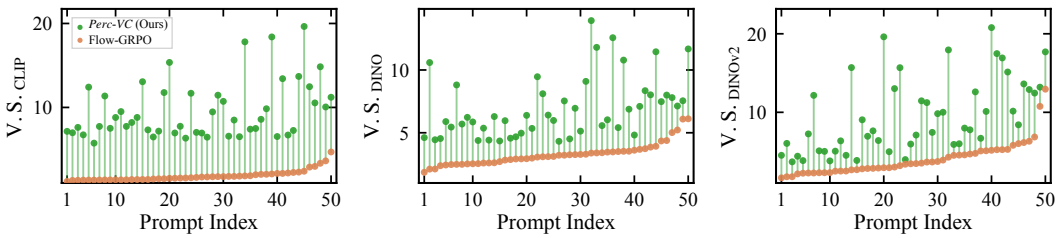

Figure 10: **Per-Prompt Vendi Scores.** For each feature space, the top-50 prompts is included, and each prompt is evaluated with 100 samples. **Left:** CLIP. **Middle:** DINO. **Right:** DINOv2.

6 Related Work

RLHF for Flow Matching. Controllable visual generation such as text-to-image [39, 67], text-to-motion [60, 61, 75, 76], and drag-based image editing [49, 50] require human preference. Thus, RLHF optimizes generative models via reward signals [38]. Building on early progress for diffusion models [5, 18, 65, 85], recent work extends RLHF to flow matching by enabling stochastic rollouts and GRPO-style optimization [39, 59, 83]. Follow-up studies improve efficiency [20, 25, 34], preference modeling [58, 89], theory [66], and reward hacking [59, 68]. Despite these advances, RLHF for flow models often suffers from diversity collapse, and its underlying mechanism remains unclear.

Entropy in LLM Reinforcement Finetuning. Entropy is widely used to characterize exploration and predict downstream gains in RLVR [12, 24, 54]. Token-level analyses further show that high-entropy tokens often correspond to key forking points and drive a disproportionate share of learning [4, 69]. Prior work thus studies entropy-aware objectives and regularization, together with stabilizers such as loss reweighting and clip-higher, to mitigate entropy collapse [11, 40, 64, 87]. In contrast, entropy in flow models can behave substantially differently, yet remains underexplored.

Entropy in Diffusion and Flow Models. Prior work studies entropy-regularized diffusion fine-tuning [63]. DIME [9] incorporates an entropy bound into critic targets, and FMER [21] combines a critic with a differentiated velocity-divergence estimator. S-MEME and FDC [13, 14] rely on optimal-control solvers requiring gradients of functional first variations, while ECFM [44] constrains entropy rates along the flow. S^2AC [45] estimates energy-based policy entropy via SVGD change of variables, and MET-SVGD [46] addresses its high-dimensional estimation limitations. In contrast, *Perc-EC* and *Perc-VC* compute perceptual entropy $\mathcal{H}_{\text{perc}}$ in a frozen encoder's feature space, rather than low-level VAE space $\mathcal{H}$, and apply it as a shaped reward on fixed rollouts, requiring no learned critic, score differentiation, or reward gradients.

7 Conclusion

In this work, we investigate diversity collapse in flow-based RLHF from the perspective of entropy mechanism, and propose a novel perceptual entropy framework with two entropy-regularization strategies that improve alignment without sacrificing diversity. First, we identify and analyze a flow-specific paradox in which perceptual diversity collapses despite the practical policy entropy surrogate remaining nearly constant. We show that this phenomenon arises from the fixed reverse-process noise schedule, whereas diversity collapse is fundamentally driven by the mode-seeking nature of policy-gradient optimization. Second, we introduce perceptual entropy, which captures diversity in a perceptual feature space while preserving key properties of standard entropy, and further propose *Perc-EC* and *Perc-VC* to maintain perceptual diversity. Extensive experiments on FLUX.dev demonstrate that our method substantially improves the quality-diversity trade-off.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (62672109, 62302093, 52441503), the Jiangsu Province Natural Science Fund (BK20230833), and the Open Research Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems under Grant E5SP060116. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, 2024. URL https://aclanthology.org/2024.acl-long.662.

[2] Isabela Albuquerque, Ira Ktena, Olivia Wiles, Ivana Kajić, Amal Rannen-Triki, Cristina Vasconcelos, and Aida Nematzadeh. Benchmarking diversity in image generation via attribute-conditional human evaluation. *arXiv preprint arXiv:2511.10547*, 2025. URL https://arxiv.org/abs/2511.10547.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI Technical Report*, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

[4] Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b6a8ca3d06ffa2e044bff8c723863ff-Abstract-Conference.html.

[5] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

[6] Black Forest Labs. FLUX, 2024. URL https://github.com/black-forest-labs/flux.

[7] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. HiDream-I1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. URL https://arxiv.org/abs/2505.22705.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

[9] Onur Celik, Zechu Li, Denis Blessing, Ge Li, Daniel Palenicek, Jan Peters, Georgia Chalvatzaki, and Gerhard Neumann. DIME: Diffusion-based maximum entropy reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 6958–6977. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/celik25a.html.

[10] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. URL https://arxiv.org/abs/2501.17811.

[11] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30377–30385, 2026. URL https://doi.org/10.1609/aaai.v40i36.40290.

[12] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025. URL https://arxiv.org/abs/2505.22617.

[13] Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Flow density control: Generative optimization beyond entropy-regularized fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pages 11056–11088, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/10715deb80d88a1ce801c313e983ff7e-Abstract-Conference.html.

[14] Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Provable maximum entropy manifold exploration via diffusion models. In *Proceedings*

of the 42nd International Conference on Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pages 12739–12761. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/de-santi25a.html.

[15] Discus0434. Aesthetic-Predictor-v2-5: SigLIP-based aesthetic score predictor, 2024. URL https://github.com/discus0434/aesthetic-predictor-v2-5.

[16] Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3029–3039, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dombrowski_Image_Generation_Diversity_Issues_and_How_to_Tame_Them_CVPR_2025_paper.html.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

[18] 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. URL https://doi.org/10.52202/075280-3497.

[19] Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

[20] Xiaolong Fu, Lichen Ma, Zipeng Guo, ShiPing Dong, Lan Yang, Tan Lit Sin, Gaojing Zhou, Yu He, Jingling Fu, Shizhe Zhou, Junshi Huang, and Jason Li. Dynamic-TreeRPO: Breaking the independent trajectory bottleneck with structured sampling. *arXiv preprint arXiv:2509.23352*, 2025. URL https://arxiv.org/abs/2509.23352.

[21] Ting Gao, Stavros Orfanoudakis, Nan Lin, Winnie Daamen, Serge P. Hoogendoorn, and Elvin Isufi. Flow matching policy optimization with mirror descent and entropy constraints. *arXiv preprint arXiv:2603.17685*, 2026. URL https://arxiv.org/abs/2603.17685.

[22] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxgknCcK7.

[23] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

[24] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

[25] 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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d75f561eaaf2cb754bc8d7e36d8af362-Abstract-Conference.html.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[27] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025. URL https://arxiv.org/abs/2501.03262.

[28] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

[29] Haozhe Ji, Cheng Lu, Yilin Niu, Pei Ke, Hongning Wang, Jun Zhu, Jie Tang, and Minlie Huang. Towards efficient exact optimization of language model alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21648–21671. PMLR, 2024. URL https://proceedings.mlr.press/v235/ji24c.html.

[30] Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*, 2025. URL https://arxiv.org/abs/2509.25133.

[31] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023. URL https://doi.org/10.52202/075280-1594.

[32] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *Technical report*, 2024. URL https://github.com/Kwai-Kolors/Kolors/.

[33] Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, and Liefeng Bo. MixGRPO: Unlocking flow-based GRPO efficiency with mixed ODE-SDE. *arXiv preprint arXiv:2507.21802*, 2025. URL https://arxiv.org/abs/2507.21802.

[34] 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. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/233d16f17f809981763db2f01b7f9603-Abstract-Conference.html.

[35] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained Chinese understanding, 2024. URL https://arxiv.org/abs/2405.08748.

[36] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-resolution text-to-3D content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lin_Magic3D_High-Resolution_Text-to-3D_Content_Creation_CVPR_2023_paper.html.

[37] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[38] Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James T. Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *ACM Computing Surveys*, 58(9):1–37, 2026. URL https://doi.org/10.1145/3796982.

[39] 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 *Advances in Neural Information Processing Systems*, volume 38, pages 40783–40818, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

[40] Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pages 17998–18031, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1a22b912945fb7c0bdd079e792b31b6f-Abstract-Conference.html.

[41] Runtao Liu, I Chieh Chen, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. AlignGuard: Scalable safety alignment for text-to-image generation. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision (ICCV), pages 17024–17034, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_AlignGuard_Scalable_Safety_Alignment_for_Text-to-Image_Generation_ICCV_2025_paper.html.

[42] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

[43] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7739–7751, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_JanusFlow_Harmonizing_Autoregression_and_Rectified_Flow_for_Unified_Multimodal_Understanding_CVPR_2025_paper.html.

[44] Chika Maduabuchi. Entropy-controlled flow matching. In *European Conference on Computer Vision (ECCV)*, pages 505–522, 2026. URL https://doi.org/10.1007/978-3-032-37419-6_29.

[45] Safa Messaoud, Billel Mokeddem, Zhenghai Xue, Linsey Pang, Bo An, Haipeng Chen, and Sanjay Chawla. S^2AC: Energy-based reinforcement learning with Stein soft actor critic. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rAHcTCMaLc.

[46] Safa Messaoud, Skander Charni, Elaa Bouazza, Ali Pourghasemi, and Halima Bensmail. Particles don't care about Z: Towards scaling entropy estimation of unnormalized densities. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61724.

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/081b08068e4733ae3e7ad019fe8d172f-Abstract-Conference.html.

[48] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

[49] Xinyu Pu, Hongsong Wang, Jie Gui, and Pan Zhou. Drag as evidence: Motion-grounded latent recomposition for drag-based editing, 2026. URL https://arxiv.org/abs/2609.36755.

[50] Xinyu Pu, Hongsong Wang, Jie Gui, and Pan Zhou. Dragging with geometry: From pixels to geometry-guided image editing. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f6775d49c75d819d275148c9902bac57-Abstract-Conference.html.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. URL https://arxiv.org/abs/2204.06125.

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. URL https://doi.org/10.1109/cvpr52688.2022.01042.

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

[55] 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. URL https://arxiv.org/abs/2402.03300.

[56] Wei Shen, Guanlin Liu, Yu Yue, Ruofei Zhu, Qingping Yang, Chao Xin, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, volume 38, pages 15283–15319, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/164687cb815daae754d33364716e65e6-Abstract-Conference.html.

[57] 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*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[58] Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. SoPo: Text-to-motion generation using semi-online preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, pages 60714–60747, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/577cd5863ec73be4e6871340be0936ae-Abstract-Conference.html.

[59] Xiaofeng Tan, Jun Liu, Yuanting Fan, Bin-Bin Gao, Xi Jiang, Xiaochen Chen, Jinlong Peng, Chengjie Wang, Hongsong Wang, and Feng Zheng. ConsistentRFT: Reducing visual hallucinations in flow-based reinforcement fine-tuning. *arXiv preprint arXiv:2602.03425*, 2026. URL https://arxiv.org/abs/2602.03425.

[60] Xiaofeng Tan, Wanjiang Weng, Haodong Lei, and Hongsong Wang. EasyTune: Efficient step-aware fine-tuning for diffusion-based motion generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/00b67df24009747e8bbed4c2c6f9c825-Abstract-Conference.html.

[61] Xiaofeng Tan, Wanjiang Weng, Hongsong Wang, Fang Zhao, Xin Geng, and Liang Wang. MotionRFT: Unified reinforcement fine-tuning for text-to-motion generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2026. URL https://doi.org/10.1109/TPAMI.2026.3741371.

[62] 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. URL https://arxiv.org/abs/2503.20314.

[63] Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024. URL https://arxiv.org/abs/2407.13734.

[64] Abdullah Vanlioglu. Entropy-guided sequence weighting for efficient exploration in RL-based LLM fine-tuning. *arXiv preprint arXiv:2503.22456*, 2025. URL https://arxiv.org/abs/2503.22456.

[65] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8238, 2024. URL https://doi.org/10.1109/cvpr52733.2024.00786.

[66] Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025. URL https://arxiv.org/abs/2509.05952.

[67] Haonan Wang, Linyu Wu, Qian Qiu, Lewei Jin, Bowen Ping, Jianghai Chen, Yiheng Du, Guangxin He, Yu Shi, Yongguang Lin, Zhuoxin Zhou, Zhanchao Zhou, Keming Wu, Rizhen Hu, Xuefei Ning, Lvfang Tao, Feiyu Hu, Xiangyan Liu, Siqi Kou, Jiarui Yao, Xiangxin Zhou, Liefeng Bo, Wenxi Zhu, and Tianyu Pang. UniRL: A Reinforcement Learning Framework for Unified Multimodal Models. https://github.com/Tencent-Hunyuan/UniRL, 2026.

[68] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5988–5998, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_GRPO-Guard_Mitigating_Implicit_Over-Optimization_in_Flow_Matching_via_Regulated_Clipping_CVPR_2026_paper.html.

[69] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pages 115452–115486, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Abstract-Conference.html.

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

[71] Xinzhou Wang, Yikai Wang, Junliang Ye, Fuchun Sun, Zhengyi Wang, Ling Wang, Pengkun Liu, Kai Sun, Xintong Wang, Wende Xie, Fangfu Liu, and Bin He. AnimatableDreamer: Text-guided non-rigid 3D model generation and reconstruction with canonical score distillation. In *European Conference on Computer Vision (ECCV)*, pages 321–339, 2024. URL https://doi.org/10.1007/978-3-031-72698-9_19.

[72] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. URL https://arxiv.org/abs/2503.05236.

[73] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AHvFDPi-FA.

[74] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *Advances in Neural Information Processing Systems*, volume 36, pages 8406–8441, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1a87980b9853e84dfb295855b425c262-Abstract-Conference.html.

[75] Wanjiang Weng, Xiaofeng Tan, Junbo Wang, Guo-Sen Xie, Pan Zhou, and Hongsong Wang. ReAlign: Text-to-motion generation via step-aware reward-guided alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13):10621–10629, Mar. 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/38035.

[76] Wanjiang Weng, Yongliang Wu, Xiaofeng Tan, Xingyu Zhu, Wenbo Zhu, and Hongsong Wang. BiMoGen: Bidirectional motion-text generation via unified masked discrete diffusion, 2026. URL https://arxiv.org/abs/2609.35407.

[77] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. URL https://doi.org/10.1007/bf00992696.

[78] 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. URL https://arxiv.org/abs/2508.02324.

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

[80] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 68578–68598. PMLR, 2025. URL https://proceedings.mlr.press/v267/xie25b.html.

[81] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45f0d179ef7e10eb7366550cd4e574ae-Abstract-Conference.html.

[82] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. URL https://doi.org/10.52202/075280-0700.

[83] 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. URL https://arxiv.org/abs/2505.07818.

[84] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. GPT-ImgEval: A comprehensive benchmark for diagnosing GPT4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025. URL https://arxiv.org/abs/2504.02782.

[85] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8941–8951, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yang_Using_Human_Feedback_to_Fine-tune_Diffusion_Models_without_Any_Reward_CVPR_2024_paper.pdf.

[86] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=AFDcYJKhND.

[87] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

[88] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, pages 57654–57689, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/537d5aa768c2d534016a4d06f87bc8fb-Abstract-Conference.html.

[89] 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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8e68ddfe22520b45b8fb8d5cbde5a21-Abstract-Conference.html.

[90] Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pages 126546–126573, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b8026fe01f5f8f7a80a1a9bf01d08cad-Abstract-Conference.html.

[91] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008. URL https://cdn.aaai.org/AAAI/2008/AAAI08-227.pdf.

Appendix

This appendix presents additional experimental results, including extended visualizations and user studies in Sec. A.1, reward variance analysis in Sec. A.2, and rule-based reward results on SD3.5-M in Sec. A.3. It further provides theoretical proofs and derivations in Sec. B, followed by implementation details of covariance-aware entropy control in Sec. C, and broader impact in Sec. D.

Contents

A Experiments **18**

A.1 Extended Visualization and User Study 18

A.2 Reward Variance Analysis on Other Perceptual Spaces 19

A.3 Rule-Based Reward Results on SD3.5-M 19

B Theoretical Proofs and Derivations **20**

B.1 Proof of Property 1 20

B.2 Proof of Latent-Space Conditional Variance Interpretation 21

B.3 Proof of Corollary 1 21

B.4 Derivation for Remark 1 23

C Covariance-Aware Entropy Control for Flow GRPO **24**

C.1 Covariance-Based Sample Identification 24

C.2 Entropy-Aware Constraint via Gradient Clipping 24

C.3 Entropy-Aware Constraint via KL Regularization 25

D Broader Impact **25**

A Experiments

A.1 Extended Visualization and User Study

Our method demonstrates strong diversity preservation across diverse prompts and domains. Fig. A1 presents an extensive collection of generated images showcasing the diversity of our approach. *Perc-EC* effectively balances quality and diversity: images maintain high aesthetic and semantic quality while exhibiting substantial visual variation within each prompt category, demonstrating that our perceptual entropy constraint successfully mitigates diversity collapse.

To further validate the effectiveness of our approach, we conduct a user study comparing our *Perc-EC* method with Flow-GRPO. The study follows a blind, randomized, side-by-side protocol: participants are shown two anonymized image sets generated by the two methods in random order and are asked to choose the better set, or indicate no clear preference, under three criteria: overall preference, diversity, and quality. We collect 100 annotations over 20 unique image-set pairs. Specifically, we recruited 20 volunteer participants from graduate students and researchers familiar with text-to-image generation. Each participant annotated five randomized image-set pairs, yielding five independent annotations per unique pair. As shown in Fig. A2, *Perc-EC* is preferred over Flow-GRPO in 64% of overall judgments and 82% of diversity judgments, with 20% and 10% no-preference responses, respectively. For image quality, participants rate the two methods comparably, with 38% preferring *Perc-EC*, 43% preferring Flow-GRPO, and 19% indicating no clear preference. The results demonstrate that our perceptual entropy constraint improves subjective diversity and overall preference while maintaining comparable image quality.

Figure A1: **Extended Diversity Visualization.** *Perc-EC* preserves visible variation across prompts and visual domains.

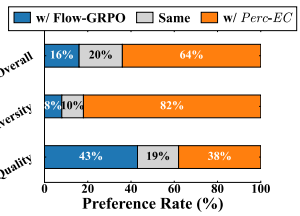

Figure A2: **User Study Preferences.**

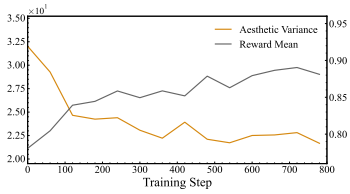

Figure A3: **Aesthetic Space Variance.**

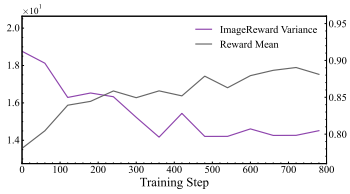

Figure A4: **ImageReward Space Variance.**

A.2 Reward Variance Analysis on Other Perceptual Spaces

To verify that our findings generalize beyond the default perceptual space, we analyze reward variance dynamics in alternative perceptual spaces defined by the Aesthetic and ImageReward encoders. As shown in Figs. A3 and A4, the same diversity-quality trade-off pattern emerges across different perceptual spaces: vanilla Flow-GRPO progressively reduces reward variance during training, reflecting the implicit drive toward mode collapse identified in our theoretical analysis. These results confirm that the entropy-variance relationship established in our framework is not an artifact of a particular feature space but a general property of policy gradient-based flow RLHF.

A.3 Rule-Based Reward Results on SD3.5-M

To complement the FLUX.dev experiments with neural rewards, we further consider the rule-based reward setting used in Flow-GRPO [39]. For completeness, Tab. A1 reproduces the GenEval comparison table from Flow-GRPO and includes our *Perc-EC* variant under the same SD3.5-M family setting.

Table A1: **GenEval Result.** Results for models other than SD3.5-M are from [84] or the corresponding original papers. Obj.: Object; Attr.: Attribute binding.

Model	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Diffusion Models							
LDM [53]	0.37	0.92	0.29	0.23	0.70	0.02	0.05
SD1.5 [53]	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD2.1 [53]	0.50	0.98	0.51	0.44	0.85	0.07	0.17
SD-XL [47]	0.55	0.98	0.74	0.39	0.85	0.15	0.23
DALLE-2 [52]	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALLE-3 [3]	0.67	0.96	0.87	0.47	0.83	0.43	0.45
Autoregressive Models							
Show-o [81]	0.53	0.95	0.52	0.49	0.82	0.11	0.28
Emu3-Gen [70]	0.54	0.98	0.71	0.34	0.81	0.17	0.21
JanusFlow [43]	0.63	0.97	0.59	0.45	0.83	0.53	0.42
Janus-Pro-7B [10]	0.80	0.99	0.89	0.59	0.90	0.79	0.66
GPT-4o [28]	0.84	0.99	0.92	0.85	0.92	0.75	0.61
Flow Matching Models							
FLUX.1 Dev [6]	0.66	0.98	0.81	0.74	0.79	0.22	0.45
SD3.5-L [17]	0.71	0.98	0.89	0.73	0.83	0.34	0.47
SANA-1.5 4.8B [80]	0.81	0.99	0.93	0.86	0.84	0.59	0.65
SD3.5-M [17]	0.63	0.98	0.78	0.50	0.81	0.24	0.52
SD3.5-M+Flow-GRPO	0.95	**1.00**	**0.99**	0.95	0.92	0.99	0.86
***+Perc-EC* (Ours)**	**0.97**	**1.00**	**0.99**	**0.97**	**0.94**	**1.00**	**0.91**

B Theoretical Proofs and Derivations

B.1 Proof of Property 1

In this section, we establish Property 1: under the reverse transition in Sec. 2, per-step conditional entropy is independent of θ. We restate the claim for convenience.

Property (Constant Per-step Entropy in Flow-based RL)**.** For the reverse-time transition defined in Eq. (4), the per-step conditional entropy at timestep t satisfies:

$$\mathcal{H}_t(p_\theta) = \frac{d}{2} - C_t.$$

where d is the dimension of $\mathbf{x}_t$, and C_t is a constant depending on timestep t.

Proof. By substituting Eqs. (4) and the log-probability definition into Eq. (9), the following key property can be derived.

We analyze the entropy term $\mathcal{H}(p_\theta) = -\mathbb{E}_{p_\theta}[\log p_\theta]$ in the gradient computation. To simplify the analysis, we directly consider sampling from p_θ.

For a continuous Gaussian transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$ at fixed timestep t, the per-step conditional entropy is

$$\mathcal{H}_t(p_\theta) = -\mathbb{E}_{\mathbf{x}_{t-1},\mathbf{x}_t}\left[\log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)\right]. \tag{A1}$$

Since $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathrm{d}t\, \mathbf{I})$, we have $\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t) + \sigma_t\sqrt{\mathrm{d}t}\cdot\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The log-probability is:

$$\log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = -\frac{\|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t, t)\|^2}{2\sigma_t^2 \mathrm{d}t} + C_t \tag{A2}$$

Substituting the sampling formula $\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t) + \sigma_t\sqrt{\mathrm{d}t}\cdot\boldsymbol{\epsilon}$ gives

$$\log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = -\frac{\|\sigma_t\sqrt{\mathrm{d}t}\cdot\boldsymbol{\epsilon}\|^2}{2\sigma_t^2 \mathrm{d}t} + C_t = -\frac{\|\boldsymbol{\epsilon}\|^2}{2} + C_t. \tag{A3}$$

Taking the expectation over $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ yields

$$\mathcal{H}_t(p_\theta) = -\mathbb{E}_{\boldsymbol{\epsilon}}\left[-\frac{\|\boldsymbol{\epsilon}\|^2}{2} + C_t\right] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{\|\boldsymbol{\epsilon}\|^2}{2}\right] - C_t = \frac{d}{2} - C_t. \tag{A4}$$

where we used $\mathbb{E}[\|\boldsymbol{\epsilon}\|^2] = d$. Here $\|\boldsymbol{\epsilon}\|^2$ follows a chi-squared distribution $\chi^2(d)$ with d degrees of freedom, so its mean is d.

Therefore, we conclude:

$$\boxed{\mathcal{H}_t(p_\theta) = \frac{d}{2} - C_t} \tag{A5}$$

This completes the proof. □

B.2 Proof of Latent-Space Conditional Variance Interpretation

In this section, we connect the per-step entropy $\mathcal{H}(p_\theta)$ in flow matching to the expected conditional variance of sampled latent variables $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$. Consistent with prior work [39, 83], $\mathbf{x}$ refers to samples in the flow model's latent space rather than images in pixel space. We use $\mathrm{Var}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ as a shorthand for the scalar $\mathrm{tr}(\mathrm{Cov}(\mathbf{x}_{t-1} \mid \mathbf{x}_t))$, i.e., the (sum of marginal) conditional variance of $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$. We restate the claim formally as a Lemma.

Lemma (Conditional Variance Interpretation of Per-Step Entropy)**.** *For the reverse-time Gaussian transition in Eq. (4), the per-step entropy in flow matching satisfies*

$$\mathcal{H}(p_\theta) = \mathbb{E}_{t,\mathbf{x}_t}\left[\frac{\mathrm{Var}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{2\sigma_t^2 \mathrm{d}t} - C_t\right],$$

where $C_t = -\frac{d}{2}\log(2\pi\sigma_t^2 \mathrm{d}t)$ is a timestep-dependent constant independent of θ.

Proof. We begin by analyzing the entropy of the conditional distribution $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)$. The derivation follows the same approach as Property 1. Recall that

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = \mathcal{N}(\mathbf{x}_{t-1};\, \mu_\theta(\mathbf{x}_t, t),\, \sigma_t^2 \mathrm{d}t \cdot \mathbf{I}), \tag{A6}$$

with the corresponding log-probability

$$\log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = -\frac{\|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t, t)\|^2}{2\sigma_t^2 \mathrm{d}t} + C_t. \tag{A7}$$

The differential entropy is then

$$\mathcal{H}(p_\theta) = -\mathbb{E}_{\mathbf{x}_{t-1},\mathbf{x}_t,t}[\log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)] = \mathbb{E}_{\mathbf{x}_{t-1},\mathbf{x}_t,t}\left[\frac{\|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t, t)\|^2}{2\sigma_t^2 \mathrm{d}t} - C_t\right]. \tag{A8}$$

Recall from Eq. (4) that $\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t) + \sigma_t\sqrt{\mathrm{d}t}\,\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so the conditional mean is

$$\bar{\mathbf{x}}_{t-1} := \mathbb{E}[\mathbf{x}_{t-1} \mid \mathbf{x}_t] = \mu_\theta(\mathbf{x}_t, t). \tag{A9}$$

Plugging this in and pulling the inner expectation through, we obtain

$$\mathcal{H}(p_\theta) = \mathbb{E}_{t,\mathbf{x}_t}\left[\frac{1}{2\sigma_t^2 \mathrm{d}t}\,\mathbb{E}_{\mathbf{x}_{t-1}|\mathbf{x}_t}\left[\|\mathbf{x}_{t-1} - \bar{\mathbf{x}}_{t-1}\|^2\right] - C_t\right]. \tag{A10}$$

By definition, the expected squared deviation from the conditional mean is the trace of the conditional covariance, which equals the (sum of marginal) conditional variance:

$$\mathbb{E}_{\mathbf{x}_{t-1}|\mathbf{x}_t}\left[\|\mathbf{x}_{t-1} - \bar{\mathbf{x}}_{t-1}\|^2\right] = \mathrm{tr}(\mathrm{Cov}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) = \mathrm{Var}(\mathbf{x}_{t-1} \mid \mathbf{x}_t). \tag{A11}$$

Substituting back and absorbing the time-step expectation gives the claimed identity:

$$\boxed{\mathcal{H}(p_\theta) = \mathbb{E}_{t,\mathbf{x}_t}\left[\frac{\mathrm{Var}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{2\sigma_t^2 \mathrm{d}t} - C_t\right].} \tag{A12}$$

This completes the proof. □

B.3 Proof of Corollary 1

We first restate the corollary to be proven, then provide the detailed proof.

Corollary (Mode-Seeking Optimization under On-Policy Policy Gradient)**.** *For policy gradient methods [77] in Eq. (2), with $p_{\theta_{\text{old}}}$ and $\mathcal{A}$ fixed within an update, the gradient satisfies:*

$$\nabla_\theta \mathcal{J}(\theta) = \nabla_\theta \Big(- D_{\text{KL}}\big(p_\theta(\cdot|c) \,\|\, p_{\mathcal{A}}(\cdot|c)\big) + D_{\text{KL}}\big(p_\theta(\cdot|c) \,\|\, p_{\theta_{\text{old}}}(\cdot|c)\big)\Big),$$

where all distributions are over trajectories $\mathbf{x}_{0:T}$, and $p_{\mathcal{A}}(\mathbf{x}_{0:T}|c) = Z_{\mathcal{A}}(c)^{-1} p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c) \exp(\mathcal{A}(\mathbf{x}_0, c))$ with $Z_{\mathcal{A}}(c) = \int p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c) \exp(\mathcal{A}(\mathbf{x}_0, c)) \, \mathrm{d}\mathbf{x}_{0:T}$.

Proof. Starting from the trajectory-level REINFORCE policy gradient in Eq. (2):

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\mathbf{x}_{0:T} \sim p_\theta(\cdot|c)} \Big[\mathcal{A}(\mathbf{x}_0, c) \sum_{t=1}^{T} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) \Big]. \tag{A13}$$

Using the trajectory factorization

$$\log p_\theta(\mathbf{x}_{0:T}|c) = \log p(\mathbf{x}_T) + \sum_{t=1}^{T} \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c), \tag{A14}$$

and taking ∇_θ on both sides, the noise prior $p(\mathbf{x}_T)$ is independent of θ so $\nabla_\theta \log p(\mathbf{x}_T) = 0$. This gives

$$\nabla_\theta \log p_\theta(\mathbf{x}_{0:T}|c) = \sum_{t=1}^{T} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c). \tag{A15}$$

Therefore:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\mathbf{x}_{0:T} \sim p_\theta} \big[\mathcal{A}(\mathbf{x}_0, c) \, \nabla_\theta \log p_\theta(\mathbf{x}_{0:T}|c) \big]. \tag{A16}$$

We rewrite $\mathcal{A}(\mathbf{x}_0, c)$ as a log-density. Note that the naive Boltzmann form $p_{\mathcal{A}} \propto \exp(\mathcal{A}(\mathbf{x}_0, c))$ on the trajectory space is not a proper density: since $\mathcal{A}$ depends only on the terminal $\mathbf{x}_0$, $\int \exp(\mathcal{A}(\mathbf{x}_0, c)) \, \mathrm{d}\mathbf{x}_{0:T}$ factors through an unbounded integral $\int 1 \, \mathrm{d}\mathbf{x}_{1:T} = \infty$ over the intermediate variables and diverges. Using $p_{\theta_{\text{old}}}$ as the base measure (as in the corollary) supplies a normalized density on $\mathbf{x}_{1:T}$ and makes $p_{\mathcal{A}}$ a valid probability distribution:

$$p_{\mathcal{A}}(\mathbf{x}_{0:T}|c) = \frac{1}{Z_{\mathcal{A}}(c)} \, p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c) \exp(\mathcal{A}(\mathbf{x}_0, c)). \tag{A17}$$

Inverting this definition gives

$$\mathcal{A}(\mathbf{x}_0, c) = \log p_{\mathcal{A}}(\mathbf{x}_{0:T}|c) - \log p_{\theta_{\text{old}}}(\mathbf{x}_{0:T}|c) + \log Z_{\mathcal{A}}(c). \tag{A18}$$

Substituting Eq. (A18) into Eq. (A16):

$$\begin{aligned} \nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{p_\theta}\big[\log p_{\mathcal{A}} \, \nabla_\theta \log p_\theta \big] - \mathbb{E}_{p_\theta}\big[\log p_{\theta_{\text{old}}} \, \nabla_\theta \log p_\theta \big] \\ + \log Z_{\mathcal{A}}(c) \cdot \mathbb{E}_{p_\theta}\big[\nabla_\theta \log p_\theta \big]. \end{aligned} \tag{A19}$$

The third term vanishes by the score-function identity $\mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta] = \nabla_\theta \, \mathbb{E}_{p_\theta}[1] = 0$.

We now apply the log-derivative trick in reverse, which requires the integrand to be θ-independent. Here $\log p_{\theta_{\text{old}}}$ is fixed by construction; for $\log p_{\mathcal{A}}$, the partition function $Z_{\mathcal{A}}(c) = \mathbb{E}_{\mathbf{x}_{0:T} \sim p_{\theta_{\text{old}}}(\cdot|c)}[\exp(\mathcal{A}(\mathbf{x}_0, c))]$ is an expectation under the frozen rollout policy and is therefore independent of θ (had it been taken under p_θ, the trick would acquire an extra $-\nabla_\theta \log Z_{\mathcal{A}}(c) \neq 0$ correction). Both terms thus satisfy

$$\mathbb{E}_{p_\theta}\big[\log p_{\mathcal{A}} \, \nabla_\theta \log p_\theta \big] = \nabla_\theta \, \mathbb{E}_{p_\theta}[\log p_{\mathcal{A}}], \qquad \mathbb{E}_{p_\theta}\big[\log p_{\theta_{\text{old}}} \, \nabla_\theta \log p_\theta \big] = \nabla_\theta \, \mathbb{E}_{p_\theta}[\log p_{\theta_{\text{old}}}]. \tag{A20}$$

Hence

$$\nabla_\theta \mathcal{J}(\theta) = \nabla_\theta \, \mathbb{E}_{p_\theta}\big[\log p_{\mathcal{A}} - \log p_{\theta_{\text{old}}} \big]. \tag{A21}$$

Adding and subtracting $\log p_\theta$ inside each expectation yields the standard cross-entropy / KL identities:

$$\mathbb{E}_{p_\theta}[\log p_{\mathcal{A}}] = -D_{\text{KL}}(p_\theta \| p_{\mathcal{A}}) - \mathcal{H}(p_\theta), \qquad \mathbb{E}_{p_\theta}[\log p_{\theta_{\text{old}}}] = -D_{\text{KL}}(p_\theta \| p_{\theta_{\text{old}}}) - \mathcal{H}(p_\theta). \tag{A22}$$

Subtracting cancels the entropy and gives

$$\mathbb{E}_{p_\theta}\big[\log p_{\mathcal{A}} - \log p_{\theta_{\text{old}}} \big] = -D_{\text{KL}}(p_\theta \| p_{\mathcal{A}}) + D_{\text{KL}}(p_\theta \| p_{\theta_{\text{old}}}). \tag{A23}$$

Combining with Eq. (A21) yields the final gradient formula:

$$\boxed{\nabla_\theta \mathcal{J}(\theta) = \nabla_\theta \Big(- D_{\text{KL}}\big(p_\theta(\cdot|c) \,\|\, p_{\mathcal{A}}(\cdot|c)\big) + D_{\text{KL}}\big(p_\theta(\cdot|c) \,\|\, p_{\theta_{\text{old}}}(\cdot|c)\big)\Big).} \tag{A24}$$

This completes the proof. □

B.4 Derivation for Remark 1

We derive the first-order relationship in Remark 1 in the main text, showing how perceptual entropy tracks the conditional feature-space variance of the reverse transition under a local smooth-encoder approximation. As in the main text, we use $\mathrm{Var}(\mathbf{z} \mid \mathbf{x})$ as a shorthand for $\mathrm{tr}(\mathrm{Cov}(\mathbf{z} \mid \mathbf{x}))$, the (sum of marginal) conditional variance of $\mathbf{z}$ given $\mathbf{x}$. We restate the claim for convenience.

Remark (*Connection to Conditional Feature-Space Variance*). Let $\mathbf{z}_{t-1} = \phi(\mathbf{x}_{t-1})$ and $\mathbf{m} = \mu_{\theta_{\text{old}}}(\mathbf{x}_t)$ as in Definition 2, with all expectations conditioned on $\mathbf{x}_t$ taken under the rollout transition $p_{\theta_{\text{old}}}$. Assume that, within a local reverse step, the perceptual encoder is approximately mean-preserving, i.e., $\mathbb{E}[\mathbf{z}_{t-1} \mid \mathbf{x}_t] \approx \phi(\mathbf{m})$. Then the perceptual entropy in Definition 2 tracks the expected conditional feature-space variance:

$$\mathcal{H}_{\text{perc}}(p_\theta) \approx \mathbb{E}_{t,\mathbf{x}_t}\left[\frac{\mathrm{Var}(\mathbf{z}_{t-1} \mid \mathbf{x}_t)}{2\sigma_t^2 \mathrm{d}t}\right] - C_t,$$

where $C_t = -\frac{d}{2}\log(2\pi\sigma_t^2\,\mathrm{d}t)$ is independent of θ.

Derivation. **Local linearization of ϕ.** Under the rollout transition, $\mathbf{x}_{t-1} \mid \mathbf{x}_t \sim \mathcal{N}(\mathbf{m},\, \sigma_t^2 \mathrm{d}t\, \mathbf{I})$ with $\mathbf{m} = \mu_{\theta_{\text{old}}}(\mathbf{x}_t)$, so $\mathbf{m}$ is exactly the conditional mean of $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$. For a frozen encoder ϕ, the second-order Taylor expansion around $\mathbf{m}$ gives

$$\phi(\mathbf{x}_{t-1}) = \phi(\mathbf{m}) + J_\phi(\mathbf{m})(\mathbf{x}_{t-1} - \mathbf{m}) + \tfrac{1}{2}(\mathbf{x}_{t-1} - \mathbf{m})^\top H_\phi(\mathbf{m})(\mathbf{x}_{t-1} - \mathbf{m}) + \mathcal{O}(\|\mathbf{x}_{t-1} - \mathbf{m}\|^3). \tag{A25}$$

Since $\mathbb{E}[\mathbf{x}_{t-1} - \mathbf{m} \mid \mathbf{x}_t] = \mathbf{0}$ holds exactly under $p_{\theta_{\text{old}}}$, the first-order term vanishes upon taking expectations. The leading residual is a second-order Jensen gap:

$$\mathbb{E}_{\mathbf{x}_{t-1}|\mathbf{x}_t}[\phi(\mathbf{x}_{t-1})] = \phi(\mathbf{m}) + \tfrac{1}{2}\sigma_t^2 \mathrm{d}t\ \mathrm{tr}(H_\phi(\mathbf{m})) + \mathcal{O}((\sigma_t^2 \mathrm{d}t)^2). \tag{A26}$$

When the encoder is locally smooth, i.e., $\sigma_t^2 \mathrm{d}t\, \|H_\phi\|$ is small, this yields the mean-preserving approximation $\mathbb{E}[\phi(\mathbf{x}_{t-1}) \mid \mathbf{x}_t] \approx \phi(\mathbf{m})$.

Perceptual entropy as a conditional residual term. Starting from Definition 2, the perceptual entropy proxy is

$$\mathcal{H}_{\text{perc}}(p_\theta) = \mathbb{E}_{t,k}\left[-\log p_\theta^{\text{perc}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, c)\right], \tag{A27}$$

where $\mathbf{z}_{t-1} = \phi(\mathbf{x}_{t-1})$ denotes samples in the perceptual space. Defining p_θ^{perc} directly as a feature-space scoring distribution omits the Jacobian of ϕ, following score distillation methods in which a frozen vision or diffusion model supplies a surrogate optimization signal in feature space [36, 48, 71, 74].

Under the Gaussian score in Definition 2, p_θ^{perc} has mean $\phi(\mathbf{m})$ and isotropic covariance $\sigma_t^2\,\mathrm{d}t \cdot \mathbf{I}$, so its log-density is

$$\log p_\theta^{\text{perc}} = -\frac{\|\mathbf{z}_{t-1} - \phi(\mathbf{m})\|^2}{2\sigma_t^2 \mathrm{d}t} + C, \qquad C = -\tfrac{d}{2}\log(2\pi\sigma_t^2 \mathrm{d}t). \tag{A28}$$

Taking the negative expectation, the perceptual entropy proxy at fixed t becomes

$$\mathcal{H}_{\text{perc}}(p_\theta \mid t) = \frac{1}{2\sigma_t^2 \mathrm{d}t}\,\mathbb{E}_{\mathbf{x}_t}\,\mathbb{E}_{\mathbf{z}_{t-1}|\mathbf{x}_t}\left[\|\mathbf{z}_{t-1} - \phi(\mathbf{m})\|^2\right] - C. \tag{A29}$$

Approximation to conditional variance. Combining the mean-preserving result $\mathbb{E}[\mathbf{z}_{t-1} \mid \mathbf{x}_t] = \phi(\mathbf{m}) + \mathcal{O}(\sigma_t^2 \mathrm{d}t)$ with the bias-variance decomposition

$$\mathbb{E}\left[\|\mathbf{z}_{t-1} - \phi(\mathbf{m})\|^2 \mid \mathbf{x}_t\right] = \mathrm{tr}(\mathrm{Cov}(\mathbf{z}_{t-1} \mid \mathbf{x}_t)) + \|\mathbb{E}[\mathbf{z}_{t-1} \mid \mathbf{x}_t] - \phi(\mathbf{m})\|^2,$$

the bias term is $\mathcal{O}((\sigma_t^2 \mathrm{d}t)^2)$ and is negligible when the encoder is locally smooth, yielding

$$\mathbb{E}_{\mathbf{z}_{t-1}|\mathbf{x}_t}\left[\|\mathbf{z}_{t-1} - \phi(\mathbf{m})\|^2\right] \approx \mathrm{tr}(\mathrm{Cov}(\mathbf{z}_{t-1} \mid \mathbf{x}_t)) = \mathrm{Var}(\mathbf{z}_{t-1} \mid \mathbf{x}_t). \tag{A30}$$

Substituting back and taking the expectation over timesteps yields

$$\boxed{\mathcal{H}_{\text{perc}}(p_\theta) \approx \mathbb{E}_{t,\mathbf{x}_t}\left[\frac{\mathrm{Var}(\mathbf{z}_{t-1} \mid \mathbf{x}_t)}{2\sigma_t^2 \mathrm{d}t}\right] - C_t,} \tag{A31}$$

which shows that perceptual entropy tracks the conditional feature-space variance of the reverse transition up to a timestep-dependent affine transformation.

For a rollout group $\{\mathbf{z}_{t-1}^k\}_{k=1}^K = \{\phi(\mathbf{x}_{t-1}^k)\}_{k=1}^K$, the corresponding empirical proxy at timestep t centers each residual at the rollout-policy transition mean $\phi(\mathbf{m}^k)$ with $\mathbf{m}^k = \mu_{\theta_{\text{old}}}(\mathbf{x}_t^k, t, c)$:

$$\widehat{\text{Var}}(t) = \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{z}_{t-1}^k - \phi(\mathbf{m}^k)\|^2, \tag{A32}$$

which gives a sample estimate of $\mathbb{E}_{\mathbf{x}_t}[\text{Var}(\mathbf{z}_{t-1} \mid \mathbf{x}_t)]$ used in the reward shaping of Sec. 4.2. □

C Covariance-Aware Entropy Control for Flow GRPO

Natural policy gradient is rarely used in post-training of generative models due to its second-order optimization cost. However, its target formulation with KL constraints shares the same spirit with TRPO and PPO. Motivated by the log-probability–advantage covariance principle discussed in prior LLM RL analyses [12], we construct a flow analogue for GRPO using reverse-transition likelihoods and derive covariance-aware constraints.

C.1 Covariance-Based Sample Identification

For a batch of N transitions $\{(\mathbf{x}_{t-1}^i, \mathbf{x}_t^i, c^i)\}_{i=1}^N$ sampled from the exploration stage, we measure how strongly the log-probability and advantage co-vary for each sample. Specifically, for each transition i, we compute the centered cross-product:

$$\text{Cov}(i) = \Big(\log p_\theta(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i) - \bar{\ell}\Big) \cdot \big(\mathcal{A}_i - \bar{\mathcal{A}}\big), \tag{A33}$$

where $\bar{\ell} = \frac{1}{N}\sum_{j=1}^N \log p_\theta(\mathbf{x}_{t-1}^j \mid \mathbf{x}_t^j, c^j)$ is the batch mean log-probability and $\bar{\mathcal{A}} = \frac{1}{N}\sum_{j=1}^N \mathcal{A}_j$ is the batch mean advantage. Here, $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c^i) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t, c^i), \sigma_t^2\,\mathrm{d}t\,\mathbf{I})$ is the conditional Gaussian policy with log-probability $\log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c^i) = -\frac{\|\mathbf{x}_{t-1}-\mu_\theta(\mathbf{x}_t,t,c^i)\|^2}{2\sigma_t^2\,\mathrm{d}t} + C_t$, and $\mathcal{A}_i$ denotes the GRPO group-normalized advantage. The empirical covariance in Eq. (A33) instantiates the analogous covariance between transition log-likelihood and advantage in the flow setting.

A large positive $\text{Cov}(i)$ indicates that transitions with high log-probability receive high advantages (and vice versa), signaling a strong positive correlation between policy likelihood and performance. Following the same covariance interpretation, such transitions are treated as entropy-reducing samples during gradient optimization. The key insight is that by identifying and selectively moderating updates on these high-covariance samples, we can control the rate of entropy collapse while maintaining reward optimization.

C.2 Entropy-Aware Constraint via Gradient Clipping

The Clip-Cov strategy leverages covariance scores to selectively suppress entropy-accelerating gradients. Given covariance bounds $\omega_{\text{low}}, \omega_{\text{high}} \in \mathbb{R}$ and a clipping ratio $r \in (0, 1)$, we randomly select transitions whose covariance falls within the specified range:

$$\mathcal{I}_{\text{clip}} \sim \text{Uniform}\Big(\{i \mid \text{Cov}(i) \in [\omega_{\text{low}}, \omega_{\text{high}}]\}, \lfloor r \cdot N \rfloor\Big). \tag{A34}$$

The Clip-Cov objective then detaches these selected transitions from gradient flow:

$$\mathcal{J}_{\texttt{Clip-Cov}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta)\,\mathbf{1}_{\{i \notin \mathcal{I}_{\text{clip}}\}}. \tag{A35}$$

By zeroing out gradients on transitions with covariance in $[\omega_{\text{low}}, \omega_{\text{high}}]$, we prevent the most aggressive entropy-reducing updates while preserving normal gradient flow on the majority of the batch. This achieves a pragmatic balance: entropy-accelerating samples are suppressed, but the policy still makes meaningful progress on reward optimization. The clipping ratio r and covariance bounds $\omega_{\text{low}}, \omega_{\text{high}}$ serve as tunable parameters to control the strength of entropy regularization. In practice, ω_{high} is typically set to a very large value (effectively $+\infty$) since we aim to suppress the most aggressive entropy-reducing updates.

Integrating Perceptual Entropy into Clip-Cov. To incorporate perceptual entropy $\mathcal{H}_{\text{perc}}$ into the Clip-Cov framework, we replace the standard covariance computation with a perceptual covariance

metric. Specifically, we compute the perceptual log-probability $\log p_{\text{perc}}(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i)$ for each generated sample, and define the perceptual covariance as:

$$\text{Cov}_{\text{perc}}(i) = \left(\log p_{\text{perc}}(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i) - \bar{\ell}_{\text{perc}}\right) \cdot \left(\mathcal{A}_i - \bar{\mathcal{A}}\right), \tag{A36}$$

where $\bar{\ell}_{\text{perc}} = \frac{1}{N}\sum_{j=1}^{N} \log p_{\text{perc}}(\mathbf{x}_{t-1}^j \mid \mathbf{x}_t^j, c^j)$ is the mean perceptual log-probability and $\bar{\mathcal{A}} = \frac{1}{N}\sum_{j=1}^{N} \mathcal{A}_j$ is the batch mean advantage. Transitions with large $\text{Cov}_{\text{perc}}(i)$ indicate samples that achieve high advantage yet receive high perceptual likelihood, signaling potential diversity collapse in perceptual space. The modified Clip-Cov objective then selects samples based on $\text{Cov}_{\text{perc}}(i)$ instead of $\text{Cov}(i)$, thereby directly targeting perceptual diversity preservation.

C.3 Entropy-Aware Constraint via KL Regularization

An alternative approach applies a KL penalty directly to high-covariance transitions. Given a threshold ratio $k \in (0, 1)$ and penalty strength $\beta > 0$, we identify the top-k proportion of transitions ranked by covariance:

$$\mathcal{I}_{\text{KL}} = \left\{ i \mid \text{Rank}(\text{Cov}(i)) \leq k \cdot N \right\}, \tag{A37}$$

where $\text{Rank}(\text{Cov}(i))$ denotes the rank in descending covariance order. For these entropy-critical transitions, we impose a KL constraint:

$$\mathcal{J}_{\texttt{KL-Cov}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) - \beta\, D_{\text{KL}}\big(p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i) \,\|\, p_\theta(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i)\big)\, \mathbf{1}_{\{i \in \mathcal{I}_{\text{KL}}\}}. \tag{A38}$$

For Gaussian policies with shared covariance $\sigma_t^2\, \mathrm{d}t$, this KL divergence reduces to:

$$D_{\text{KL}}\big(p_{\theta_{\text{old}}} \| p_\theta\big) = \frac{\|\mu_{\theta_{\text{old}}}^i - \mu_\theta^i\|^2}{2\sigma_t^2\, \mathrm{d}t}. \tag{A39}$$

The KL-Cov approach penalizes large policy changes on entropy-critical samples, thereby constraining the magnitude of entropy-reducing updates. The penalty coefficient β governs the trade-off between reward optimization and entropy preservation. In contrast to Clip-Cov, which discretely suppresses gradients, KL-Cov applies a continuous penalty that allows some flexibility while still dampening aggressive updates. Together, these two strategies provide complementary mechanisms to moderate entropy decrease by directly targeting samples identified as entropy accelerators through the flow covariance criterion.

Integrating Perceptual Entropy into KL-Cov. To integrate perceptual entropy into the KL-Cov framework, we replace the sample ranking criterion with perceptual covariance. Instead of ranking by $\text{Cov}(i)$, we identify the top-k proportion of transitions ranked by perceptual covariance:

$$\mathcal{I}_{\text{KL-perc}} = \left\{ i \mid \text{Rank}(\text{Cov}_{\text{perc}}(i)) \leq k \cdot N \right\}, \tag{A40}$$

where $\text{Rank}(\text{Cov}_{\text{perc}}(i))$ denotes the rank in descending perceptual covariance order. The modified objective applies the same KL penalty to these selected transitions:

$$\mathcal{J}_{\texttt{KL-Cov-perc}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) - \beta\, D_{\text{KL}}\big(p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i) \,\|\, p_\theta(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c^i)\big)\, \mathbf{1}_{\{i \in \mathcal{I}_{\text{KL-perc}}\}}. \tag{A41}$$

By selecting samples based on perceptual covariance, the KL penalty is applied to transitions that most strongly contribute to perceptual entropy decrease, thereby directly targeting diversity preservation in perceptual space.

D Broader Impact

Positive impacts. This work improves diversity preservation in RLHF-aligned text-to-image generation, enabling models to produce a broader range of high-quality outputs for a given prompt. This benefits creative applications such as design assistance, content creation, and personalized media generation, where coverage across diverse styles and contexts is essential.

Negative impacts. Improved image generation quality and diversity may lower the barrier to producing misleading or harmful visual content, including deepfakes and synthetic disinformation. Moreover, while pretrained visual generative models typically undergo strict data safety auditing to suppress violent or pornographic outputs, visual post-training and diversity-promoting optimization can weaken these built-in safeguards and reintroduce harmful visual content. Our method is a fine-tuning technique applied on top of existing base models (e.g., FLUX.dev) that carry their own usage policies, and practical deployments should pair reward and diversity optimization with dedicated safety filters.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three main contributions: (1) theoretical and empirical analysis of the constant-policy-entropy paradox in flow-based RLHF, (2) perceptual entropy as a diversity measure, and (3) two perceptual-entropy strategies, *Perc-EC* and *Perc-VC*. These claims are supported by Secs. 3–5 across FLUX.dev and SD3.5-M, neural and rule-based rewards, and two perceptual spaces. Sec. 5 additionally evaluates generalization to DDPO and text-to-video generation on Wan2.1 using UniRL, while App. A.1 reports a blind user study.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Negative impacts are discussed in App. D. While pretrained visual generative models typically undergo strict data safety auditing to suppress violent or pornographic outputs, visual post-training and diversity-promoting optimization can weaken these built-in safeguards and reintroduce harmful visual content.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Property 1, Corollary 1, Remark 1, and the supporting lemmas state their assumptions explicitly. Complete proofs and derivations are provided in App. B.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 5 specifies the Flow-GRPO-based training setup, FLUX.dev with PickScore, Pick-a-Pic training prompts, HPD quality evaluation, and diversity evaluation using 100 samples for each of 85 PartiPrompts. It reports the quality and Vendi-score metrics, the $\lambda \in \{0.03, 0.05, 0.10\}$ search space, baseline hyperparameters, and GPU counts. The same section describes the DDPO and Wan2.1/UniRL extensions. App. A.3 provides the SD3.5-M/GenEval setting, and App. C details the covariance-aware baselines.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The project page provides public access to the implementations of *Perc-EC* and *Perc-VC*. The datasets and existing frameworks used in the paper, including Pick-a-Pic, HPD, PartiPrompts, GenEval, Flow-GRPO, DDPO, Wan2.1, and UniRL, are also publicly accessible.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Sec. 5 specifies the training and evaluation datasets, reward and perceptual encoders, $\lambda \in \{0.03, 0.05, 0.10\}$, sampling protocol, quality and diversity metrics, min–max-normalized aggregate scores, and baseline hyperparameters (entropy weight, covariance rate, clipping threshold, and KL weight). Remaining image-training hyperparameters follow the public Flow-GRPO setup. App. A.3 and App. C provide details for SD3.5-M and covariance-aware methods.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports quantitative comparisons across multiple quality and diversity metrics. App. A.1 further reports preference rates over 100 judgments on 20 unique image-set pairs, with five independent judgments per pair.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 5 reports that image experiments use 8 GPUs and video experiments use 96 GPUs.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The user study follows a blind, randomized protocol, and participants were not exposed to harmful content.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: App. D discusses positive applications in diversity-preserving creative tools and negative impacts including deepfakes, synthetic disinformation, and the possibility that visual post-training weakens safety safeguards and reintroduces harmful content.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a fine-tuning method built on top of FLUX.dev, which already has its own usage policies. The paper does not release a new pre-trained model or dataset, and the released code follows the license of the base model.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the models, reward and perceptual encoders, datasets, and training frameworks used in the experiments, including FLUX.dev, SD3.5-M, Wan2.1, PickScore, VideoPickScore, DINO, CLIP, ImageReward, HPSv2.1, Aesthetic Predictor v2.5, Unified Reward, Pick-a-Pic, HPD, PartiPrompts, GenEval, Flow-GRPO, DDPO, and UniRL. Use of these existing assets follows their respective terms.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets or pre-trained model weights. The code implementing *Perc-EC* and *Perc-VC* is publicly available through the project page.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: App. A.1 describes the blind, randomized, side-by-side protocol, the three evaluation criteria, 20 image-set pairs, 20 volunteer participants, and 100 total judgments.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study is a minimal-risk visual preference task involving volunteers, with no sensitive personal data or harmful content; formal IRB review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the methodology. The proposed methods are based on flow-matching models and perceptual encoders. Any LLM use was limited to writing, editing, formatting, and typesetting.