NEURAL INFORMATION PROCESSING SYSTEMS

SPOTLIGHT 2026

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

Project Page · Paper · arXiv · Training Log

Introduction

Problem. RLHF improves preference alignment but can collapse visual diversity, limiting exploration and perceptual coverage.



Flow-GRPO Perc-EC (Ours)

Why direct adaptation fails. Unlike token-level entropy in LLMs, native flow-policy entropy is fixed by the noise schedule. Directly transferring entropy constraints cannot track perceptual collapse.

Contributions

1. We explain why policy entropy stays constant during RLHF and why mode-seeking updates still cause perceptual diversity collapse.

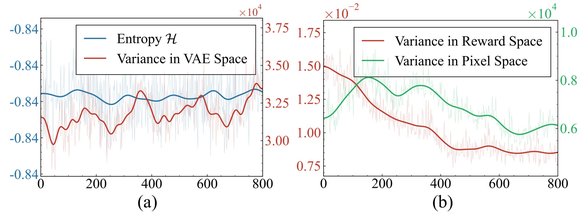
2. We introduce perceptual entropy and an empirical entropy–reward relationship, and propose Perc-EC and Perc-VC to preserve diversity.

3. We demonstrate gains on FLUX.dev and SD3.5-M across neural/rule-based rewards and PickScore and DINO perceptual encoders.

Constant Policy Entropy

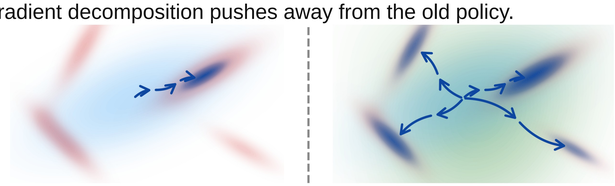
The noise schedule fixes the transition covariance. Model updates change the Gaussian mean, not the analytical policy entropy.

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



Diversity Collapse

Mode-seeking updates favor high-reward regions. The old-policy KL term in the gradient decomposition pushes away from the old policy.



Method: Perceptual Entropy

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.

Perceptual probability and entropy

Noised samples and old-policy means are first decoded into pixels, then mapped into perceptual space by a perceptual encoder.

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

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

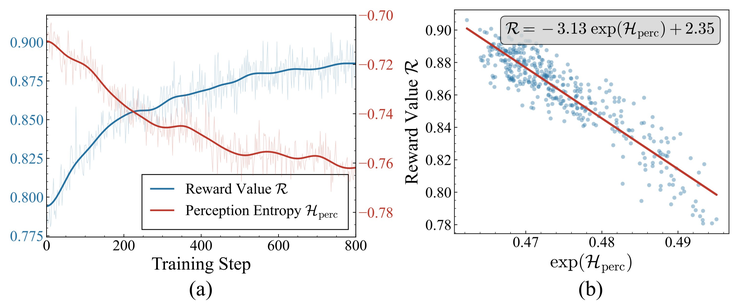
Connection to conditional feature-space variance

Within a local reverse step, assume that the perceptual encoder is approximately mean-preserving. Then perceptual entropy tracks conditional feature-space dispersion:

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

Empirical entropy mechanism

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



Reward and perceptual entropy follow this empirical relationship during training; a and b are fitted coefficients.

Perceptual Entropy Constraint (Perc-EC)

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

Directly promote perceptual dispersion alongside preference reward. Larger regularization weights encourage broader perceptual coverage.

Perceptual VAE Constraint (Perc-VC)

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

Anchor the perceptual score distribution to the diversity-preserving generation-space reference, instead of unconstrained entropy maximization.

Method: Perceptual Reward Shaping

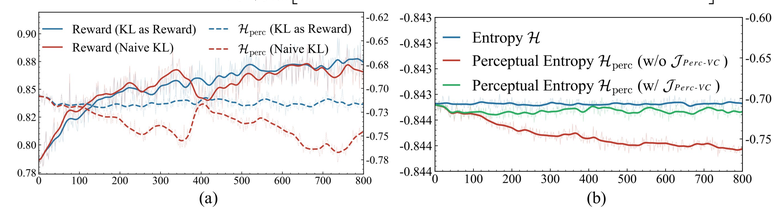
Optimizing independent auxiliary losses introduces competing gradients and instability. We instead incorporate perceptual constraints directly into the reward signal.

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

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

Task and perceptual rewards are jointly normalized within the same rollout group, followed by the standard GRPO clipped surrogate.

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

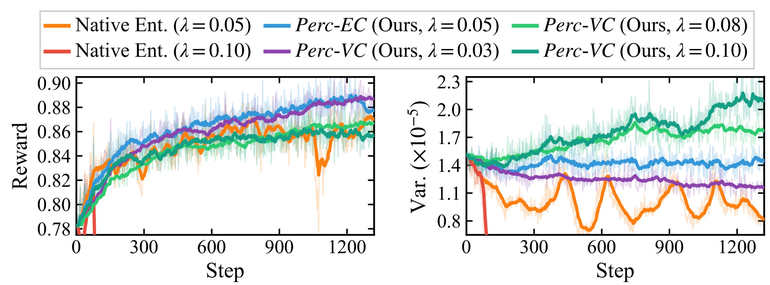


Experiments

Entropy regularization comparison

Method	Perceptual Space	Quality↑ P.S.	Aes.v2.5	I.R.	CLIP	HPSv2.1	U.R.	Norm.Avg	Diversity↑ DINO	CLIP	V.S.$_{\text{DINO}}$	V.S.$_{\text{CLIP}}$	V.S.$_{\text{IR}}$	V.S.$_{\text{PS}}$	Norm.Avg	Overall
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	0.57	0.43	3.17	2.02	1.16	1.11	0.111	0.361
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**

Generalization to DDPO



DDPO: reward (left) and within-group perceptual variance (right).