GDPO-SR: Group Direct Preference Optimization for One-Step Generative Image Super-Resolution

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, and Lei Zhang

The Hong Kong Polytechnic University, OPPO Research Institute

CVPR 2026

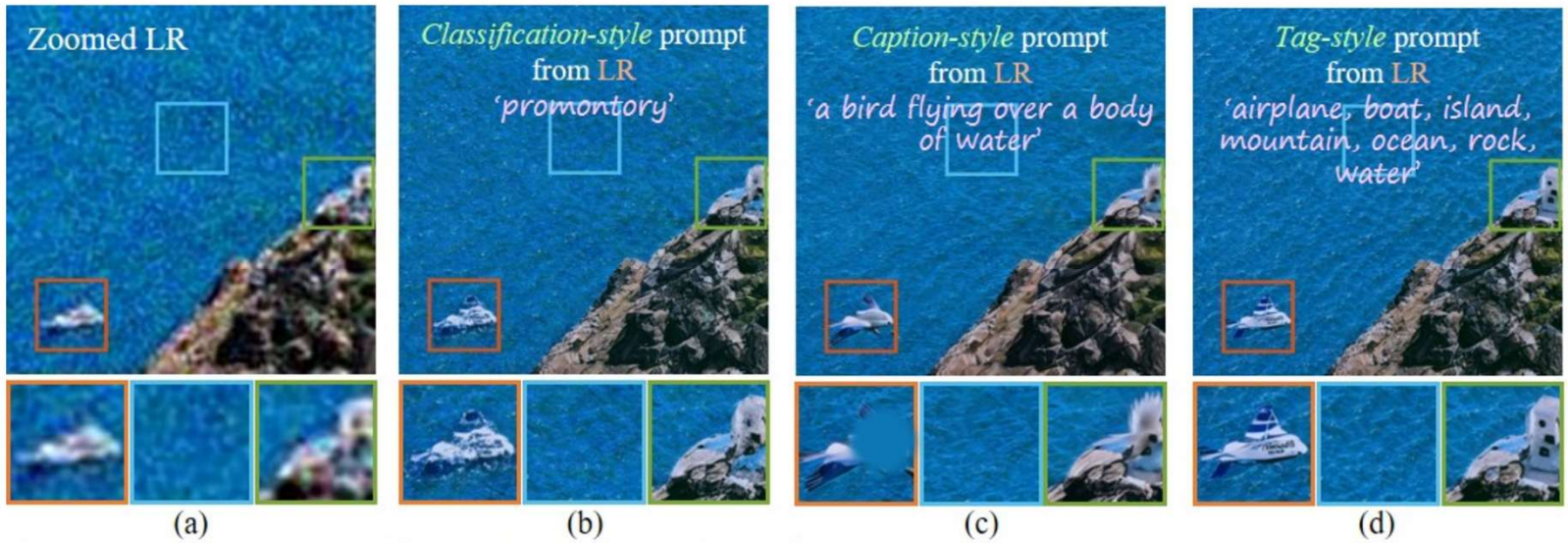
Presenter: Jinyi Luo

2026.03.29

Background

Settings and Issues of Image Super-resolution：

- Recover high-resolution images from low-resolution input with complex degradations
- **Ill-posed problem:** even with perfect training supervision, lost details cannot be deterministically recovered



Background

Stochasticity of Diffusion-based SR：

- Multi-step Diffusion SR Solves SDE equations：

$$dx = \left[-\tfrac{1}{2}\beta(t)x - \beta(t)\nabla_x \log p_t(x \mid y)\right] dt + \sqrt{\beta(t)}\, d\bar{\mathbf{w}}$$

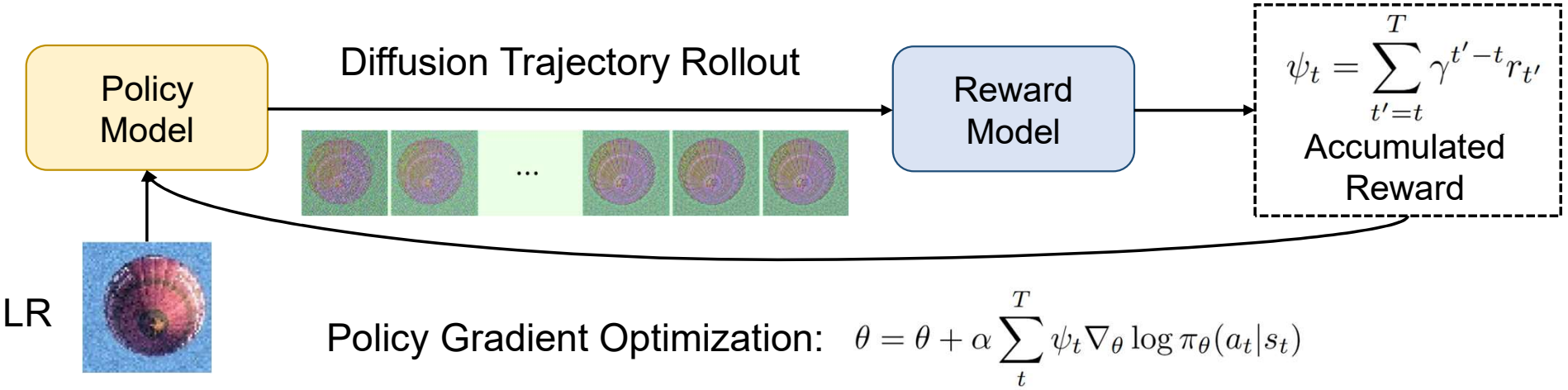
- Step-wise stochasticity: high variability in SR details
- Exploit randomness: bias details toward human preference

Rollout 1: fake semantics

Rollout 2: fake texture

Rollout 3: fits human preference

Reinforcement Learning to Leverage Stochasticity：



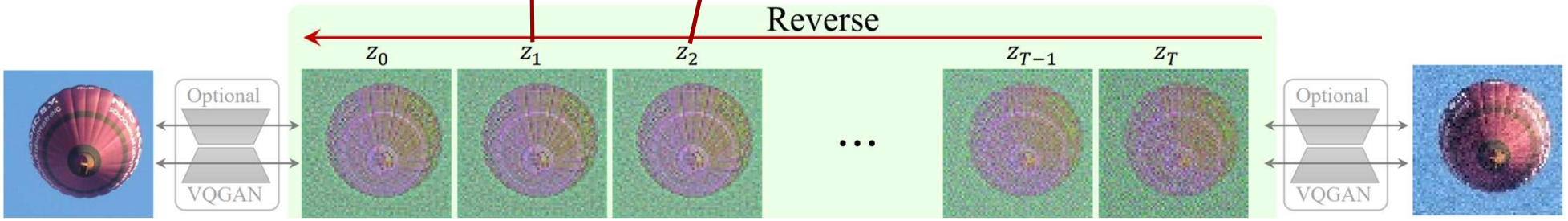
REINFORCE: $\psi_t \nabla_\theta \log \pi_\theta(a_t|s_t)$

PPO: with Advantage Sampling $\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)} A^{\theta'} \nabla \log p_\theta(a_t^n|s_t^n)$

Problem of Classical RL for Diffusion Process：

- REINFORCE & PPO relies on per-timestep reward aggregation
- **Per-timestep image quality assessment is hard** for diffusion process

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1},$$
$$\text{where} \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$



北京大学 PEKING UNIVERSITY

Solution：Directly Optimize Policy with Preference Data (DPO)

- Construct dataset with paired data of High / Low preference
- Encourage policies with higher preference results and suppress lower ones：

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

Positive　　Negative

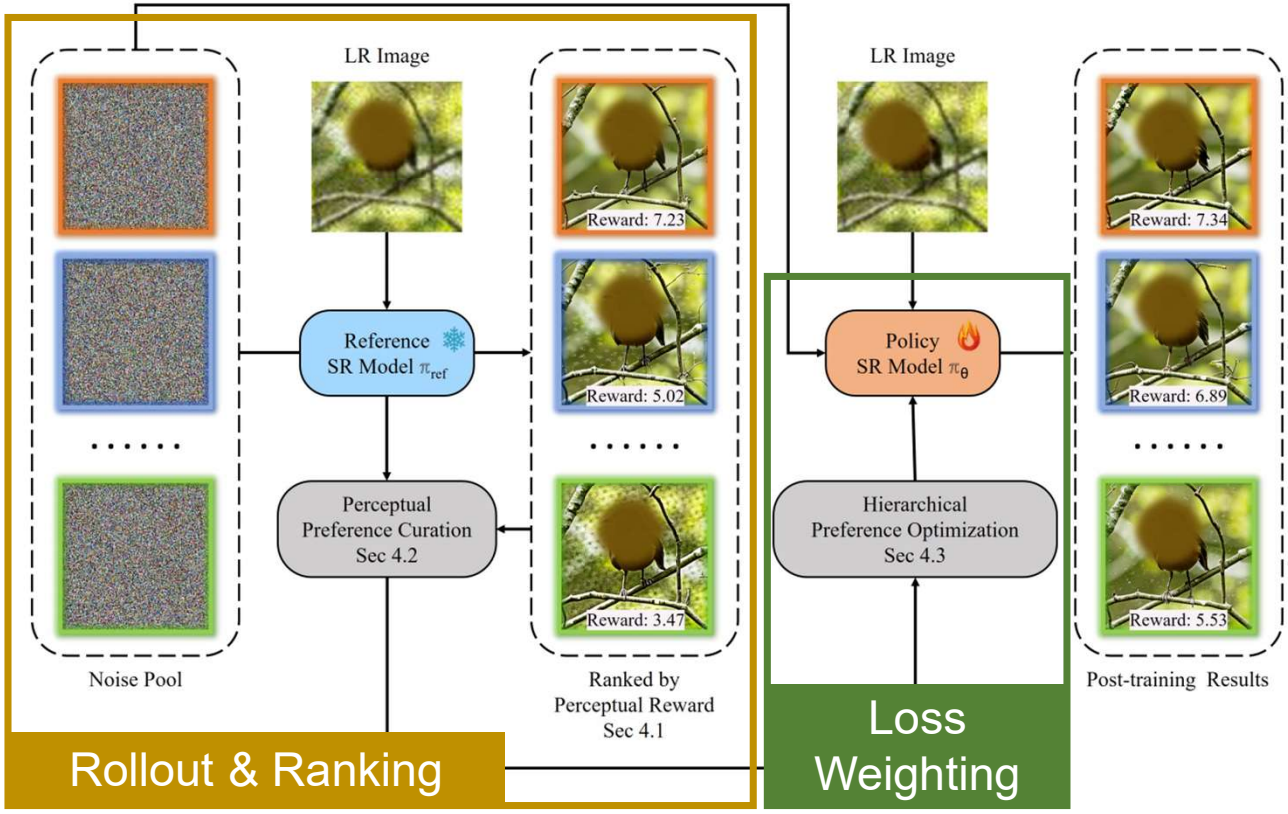
- Convert to Diffusion-DPO：

$$L_{DPO} = -\mathbb{E}_{(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim D,t\sim\mathcal{U}(0,T),\boldsymbol{x}_t^w\sim q(\boldsymbol{x}_t^w|\boldsymbol{x}_0^w),\boldsymbol{x}_t^l\sim q(\boldsymbol{x}_t^l|\boldsymbol{x}_0^l)}\log\sigma(-\beta($$
$$\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w,t)\|_2^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w,t)\|_2^2 - (\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l,t)\|_2^2 - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l,t)\|_2^2)))$$

Positive　　Negative

Wallace *et al.*, Diffusion Model Alignment Using Direct Preference Optimization, in CVPR 2024.

DPO for SR Example: DP²O-SR



- Rollout N samples
- Use IQA metrics to rank those samples, construct N^2 winner-loser pairs
- Optimize model through:

$$\mathcal{L}_{HPO} = \sum_{(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l)} w \cdot \ell(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l; \theta)$$

Wu *et al.*, DP2O-SR: Direct Perceptual Preference Optimization for Real-World Image Super-Resolution, in NeurIPS 2025.

DPO for SR Example: DP²O-SR

- Different sample pairs contribute unequally to optimization
  - **Large preference gaps** between positive and negative samples

  ——stronger optimization signals

$$w_{\text{intra}}(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l) = |R_w - R_l| + (1 - \mu_{\text{gap}})$$

  - **Greater rollout diversity** under the same LR condition

  ——more effective exploration

$$w_{\text{inter}}(g) = \sigma_g + (1 - \mu_\sigma)$$

Wu *et al*., DP2O-SR: Direct Perceptual Preference Optimization for Real-World Image Super-Resolution, in NeurIPS 2025.

Beyond Binary Optimization: GRPO

- Group Preference Policy Optimization
- One optimization step determined by group rollout results

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}$$

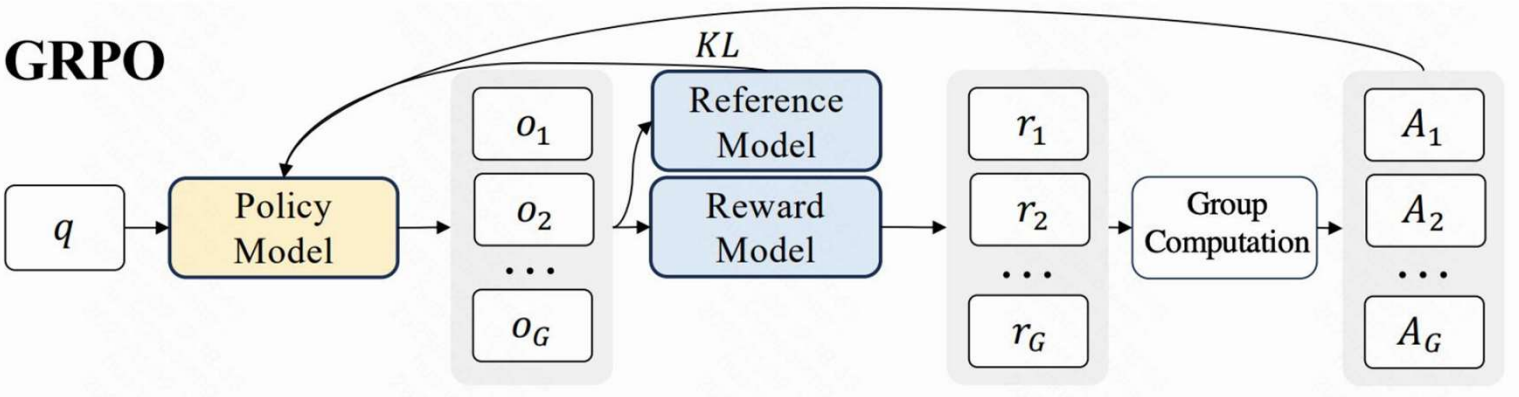
- Relative advantage of intra-group reward as state value

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

Shao *et al*., DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, arXiv 2024.

Beyond Binary Optimization: GRPO

- Group Preference Policy Optimization
- One optimization step determined by group rollout results



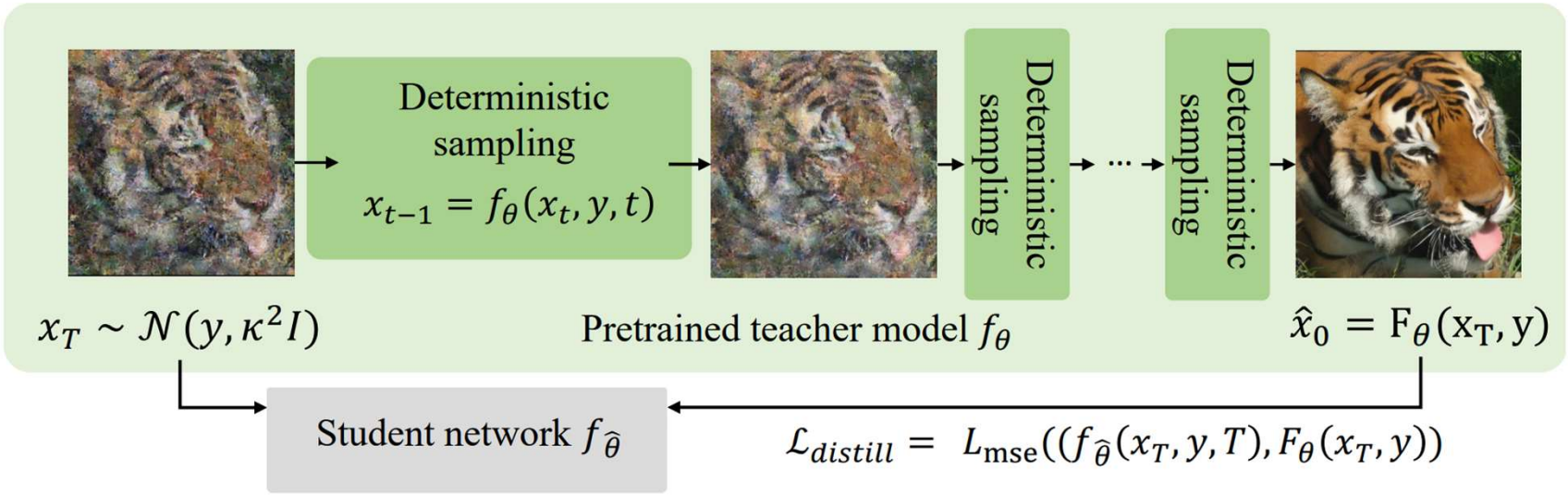
- Relative advantage of intra-group reward as state value

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

Shao *et al.*, DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, arXiv 2024.

One-step Diffusion SR: Possible Faster Rollout

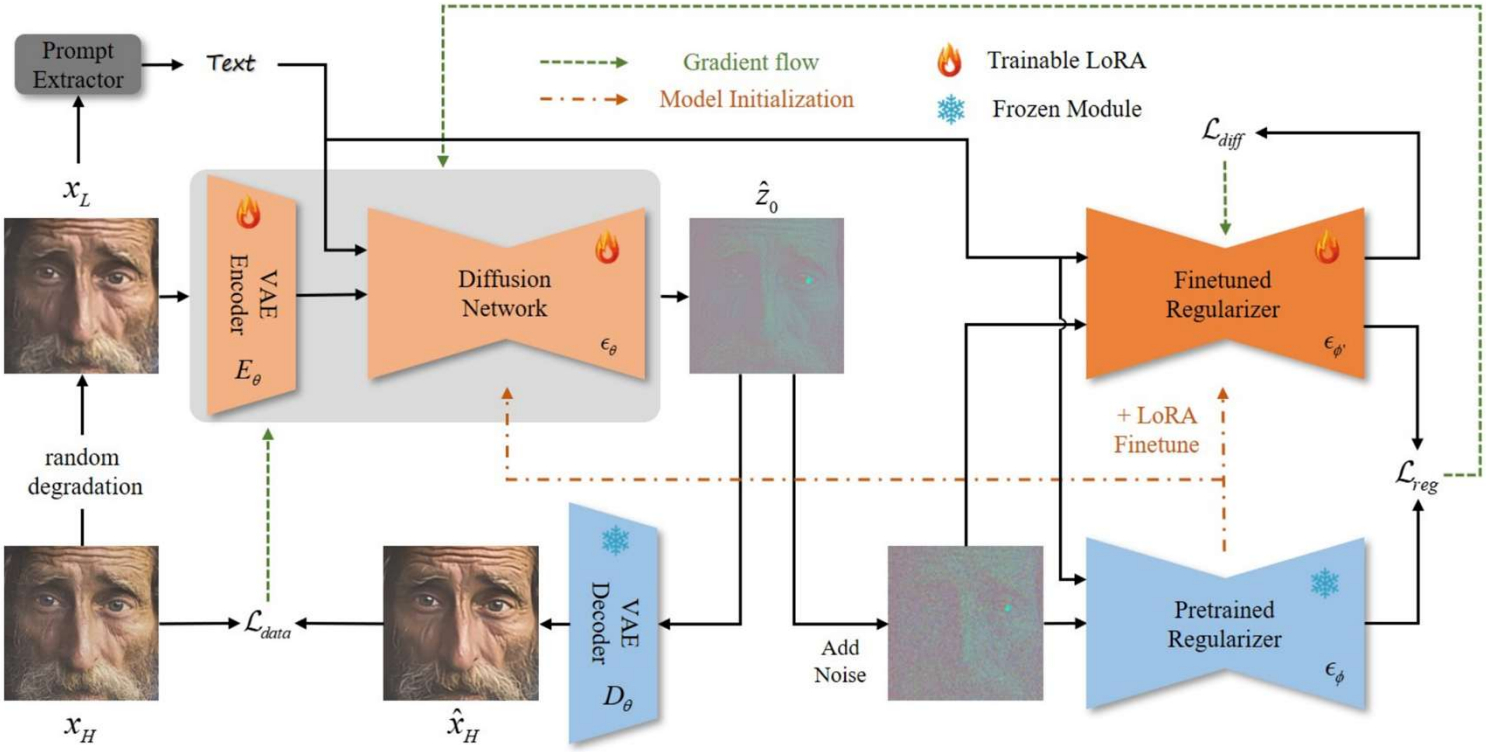
- First replace stochastic sampling with deterministic sampling
- Then distil a student model to learn the deterministic mapping



Wang *et al.*, SinSR: Diffusion-Based Image Super-Resolution in a Single Step, in CVPR 2024.

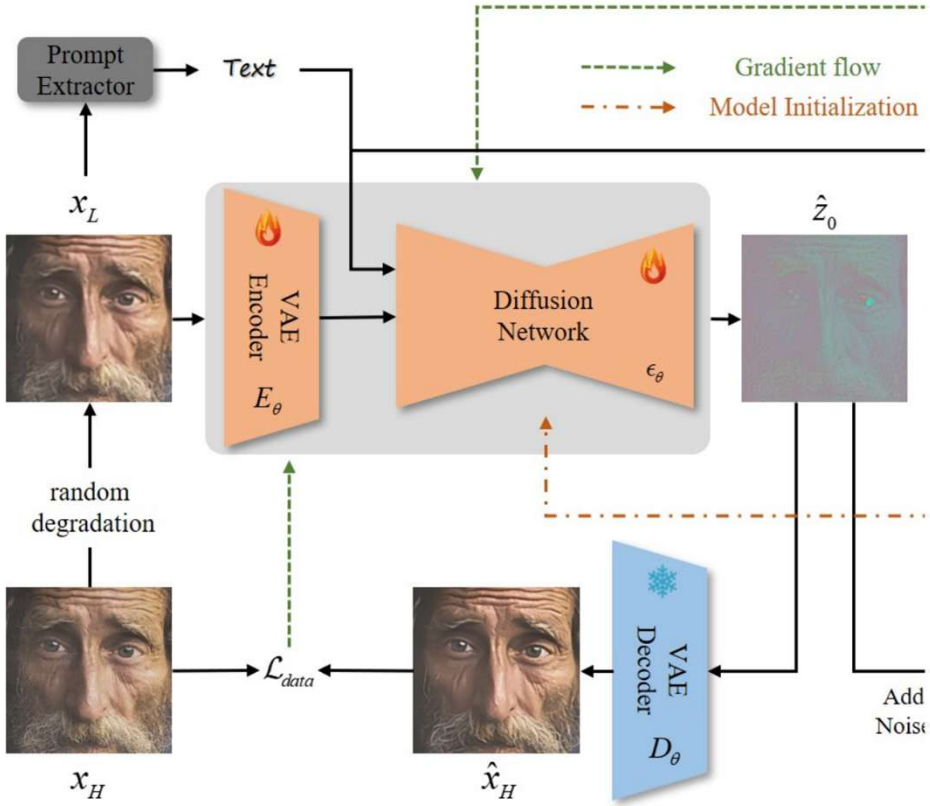
Train a One-step SR model without Teacher Model Supervision：

- SR reconstruction + Diffusion regularization



Wu *et al.*, One-Step Effective Diffusion Network for Real-World Image Super-Resolution, in NeurIPS 2024.

Train a One-step SR model without Teacher Model Supervision



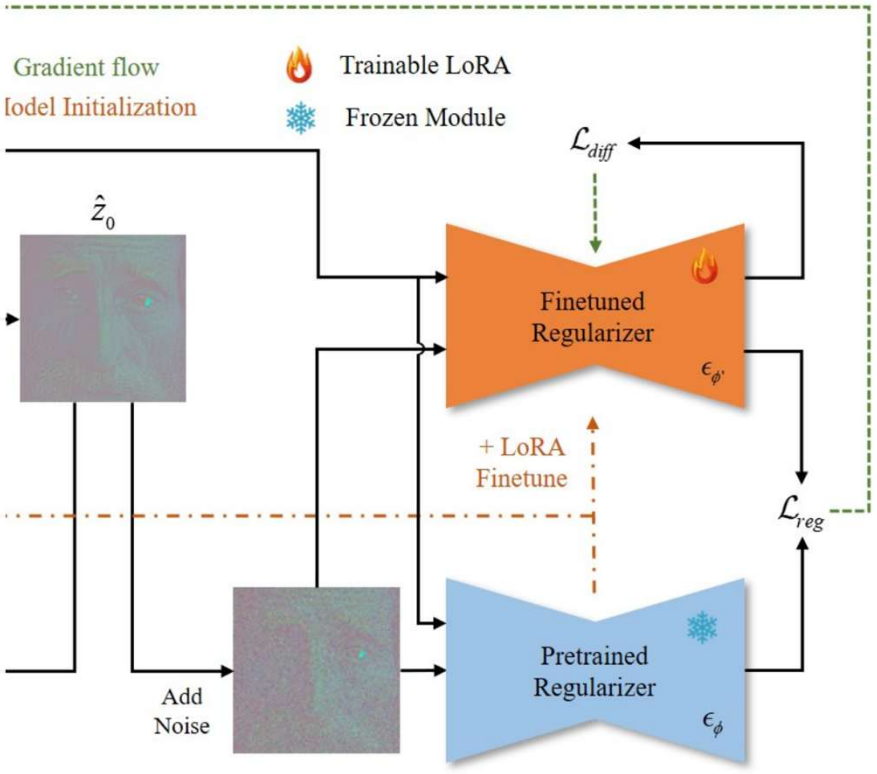
- Directly encode LR as Diffusion starting point
- Train for direct LR-HR latent mapping
- Compute L2 loss as data reconstruction supervision

$$\mathcal{L}_{\text{MSE}}\left(G_\theta(\boldsymbol{x}_L), \boldsymbol{x}_H\right) + \lambda_1 \mathcal{L}_{\text{LPIPS}}\left(G_\theta(\boldsymbol{x}_L), \boldsymbol{x}_H\right)$$

Wu *et al.*, One-Step Effective Diffusion Network for Real-World Image Super-Resolution, in NeurIPS 2024.

Train a One-step SR model without Teacher Model Supervision

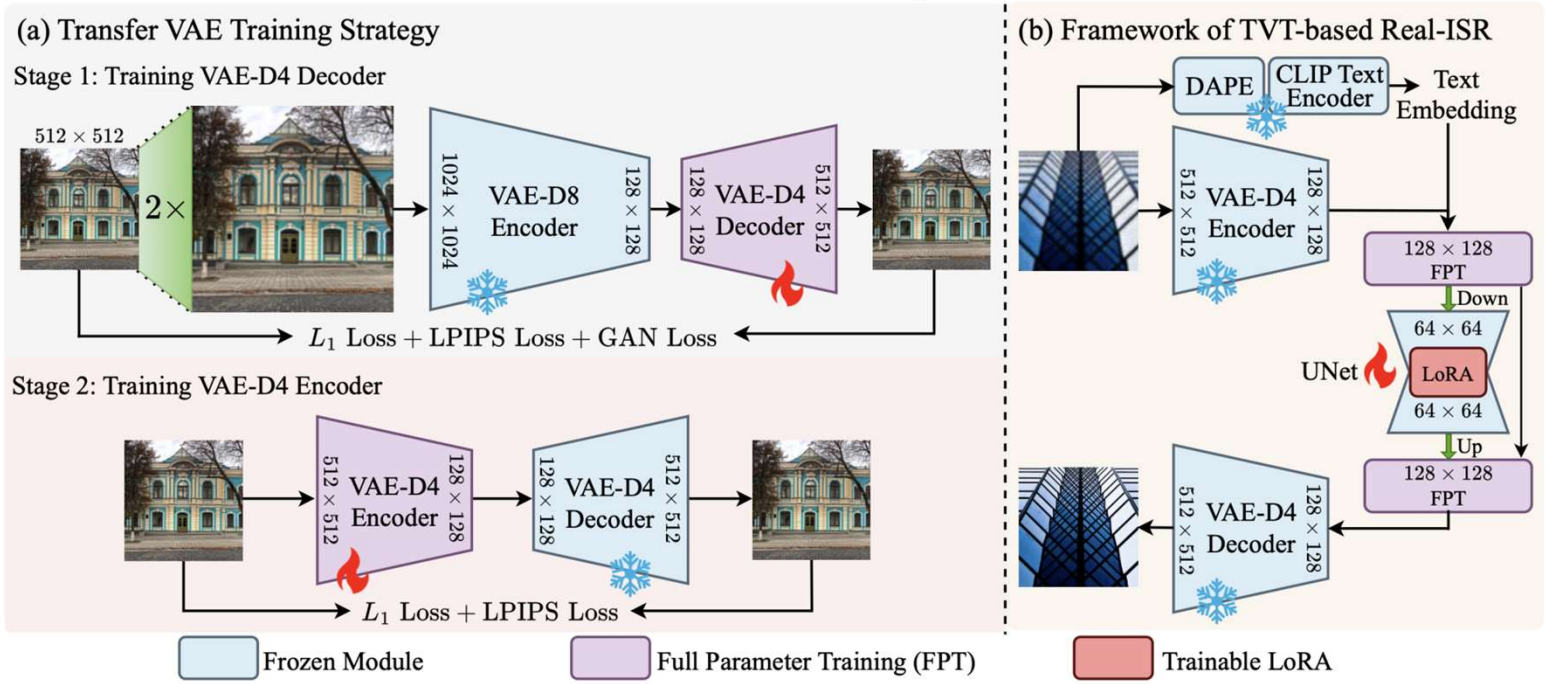
- Regularize the one-step latent to align with the diffusion model distribution
- Use two regularizers to judge distribution shift
- If the **fake (learned) score differs from the real score**, then the input domain is distorted



Wu *et al.*, One-Step Effective Diffusion Network for Real-World Image Super-Resolution, in NeurIPS 2024.

One-step Diffusion SR without Extra Regularization

- Learn a VAE with **lower downsample-rate** for better detail preservation
- LoRA-finetune a Diffusion model for one-step SR

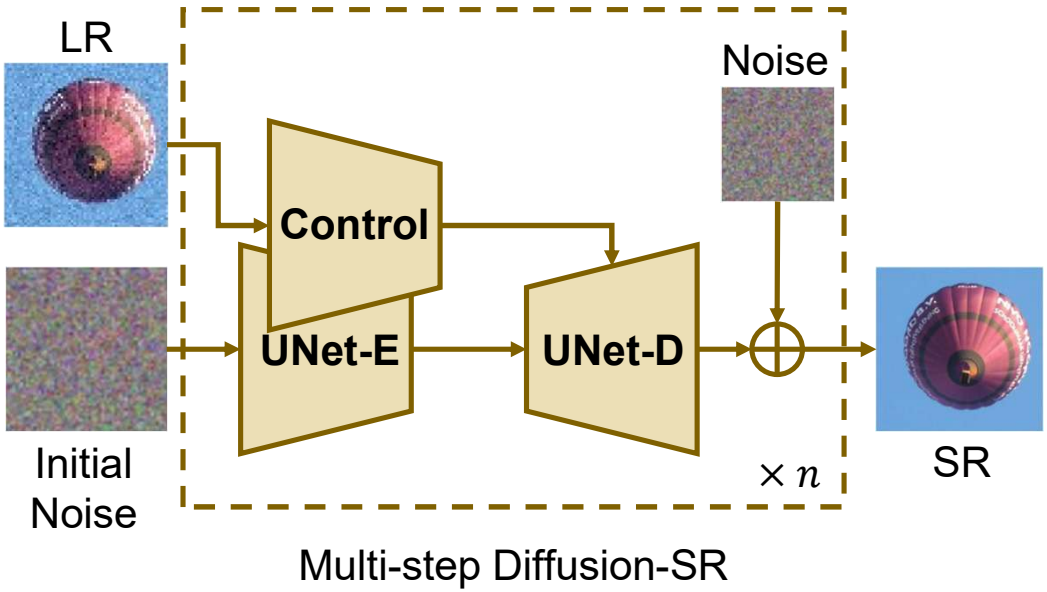


Yi *et al.*, Fine-structure Preserved Real-world Image Super-resolution via Transfer VAE Training, in ICCV 2025.

Stochasticity Problem of Applying GRPO to One-step Diffusion:

- We leverage stochasticity in model sampling for preference exploration

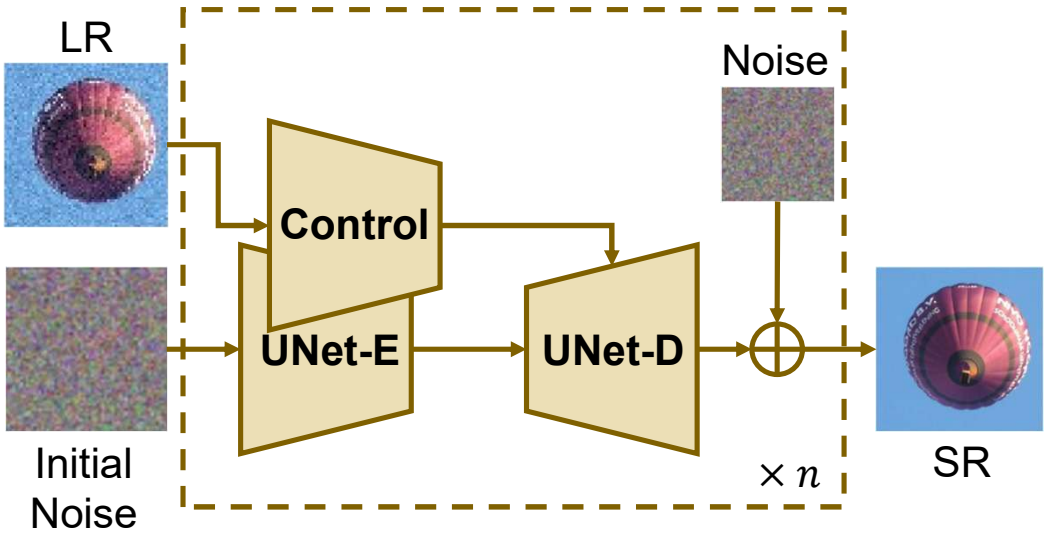
$$z_{t-1} = \frac{z_t - \sqrt{\beta_t}\,\mathrm{UNet}(z_t, c_t, t)}{\sqrt{\alpha_t}} + \sigma_t\,\epsilon$$



Stochasticity Problem of Applying GRPO to One-step Diffusion:

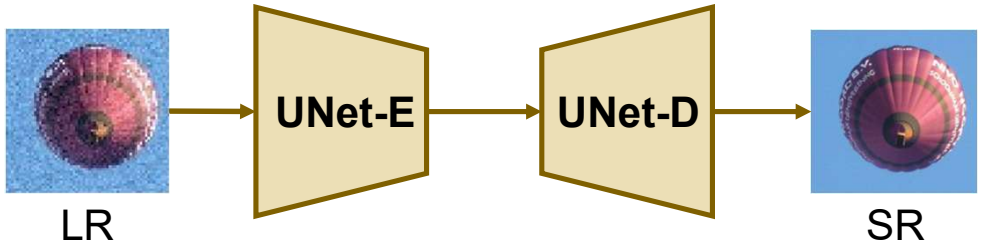
- We leverage stochasticity in model sampling for preference exploration
- However, single-step diffusion **does not have sampling stochasticity**

$$z_{t-1} = \frac{z_t - \sqrt{\beta_t}\,\mathrm{UNet}(z_t, c_t, t)}{\sqrt{\alpha_t}} + \sigma_t\,\epsilon$$



Multi-step Diffusion-SR

$$z_{SR} = \frac{z_{LR} - \sqrt{\beta_t}\mathrm{UNet}(z_{LR}, c_t, t)}{\sqrt{\alpha_t}}$$



One-step Diffusion-SR

Summary: Key Problems of Designing RL-based One-step SR models:

1. **How to inject stochasticity?**

2. **What advantage function fits better for SR task?**

Method

北京大学 PEKING UNIVERSITY

Stochasticity Injection | **Advantage Design** | **Policy Optimization**

- Stochasticity Injection in Deterministic Sampling：
- Flow-GRPO: ODE to SDE via **adding stochastic noise**

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{v}_t \mathrm{d}t$$

$$\downarrow$$

$$\mathrm{d}\boldsymbol{x}_t = \left(\boldsymbol{v}_t(\boldsymbol{x}_t) - \frac{\sigma_t^2}{2} \nabla \log p_t(\boldsymbol{x}_t) \right) \mathrm{d}t + \sigma_t \mathrm{d}\boldsymbol{w}$$

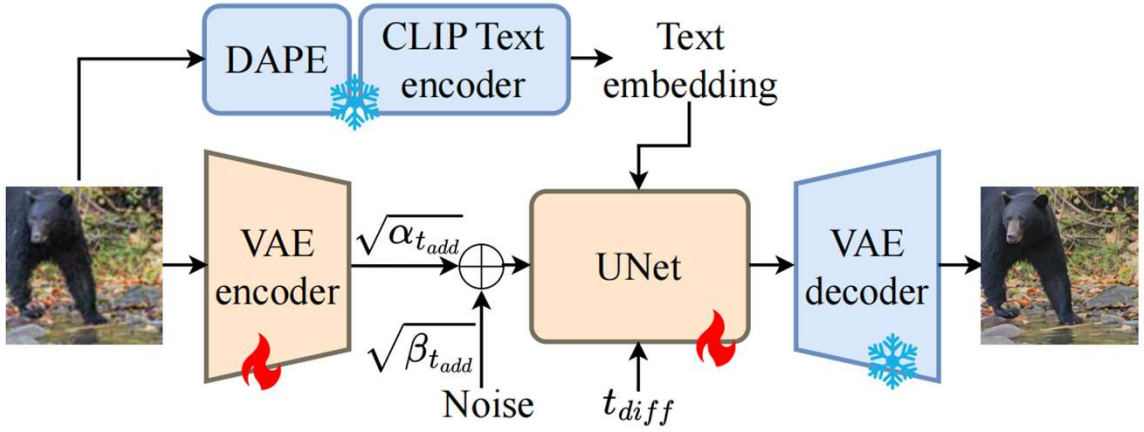
Liu *et al*., Flow-GRPO: Training Flow Matching Models via Online RL, arXiv 2025.

Method

北京大学 PEKING UNIVERSITY

Stochasticity Injection | **Advantage Design** | **Policy Optimization**

- Stochasticity Injection in Deterministic Sampling：
- Flow-GRPO: ODE to SDE via adding stochastic noise
- **Similarly, add noise to LR latent before one-step inference**



Method

北京大学 PEKING UNIVERSITY

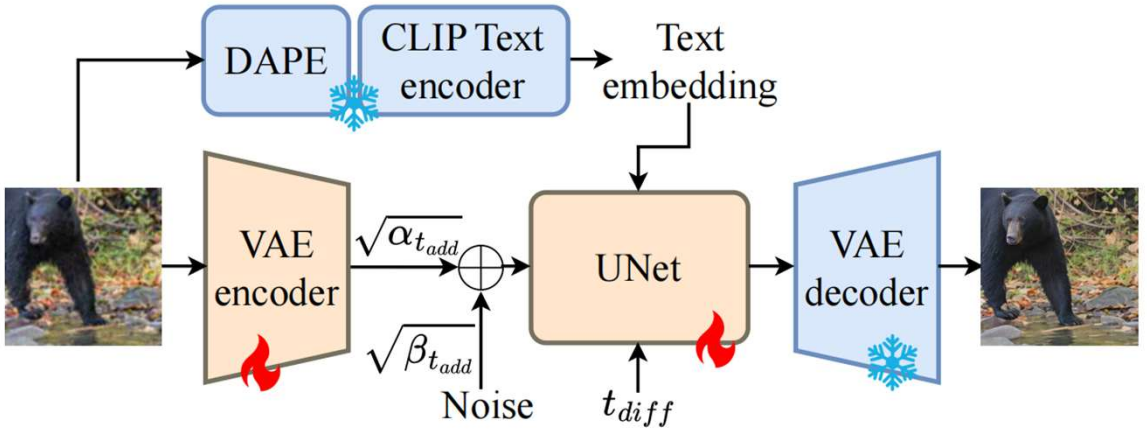
Stochasticity Injection | **Advantage Design** | **Policy Optimization**

- Similarly, add noise to LR latent before one-step inference

$$\tilde{z} = \sqrt{\alpha_{t_{add}}} z_{LR} + \sqrt{\beta_{t_{add}}} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

$$z_{SR} = \frac{(\tilde{z} - \sqrt{\beta_{t_{diff}}}\ \mathrm{UNet}(\tilde{z}, c_t, t_{diff}))}{\sqrt{\alpha_{t_{diff}}}}$$

- Pretrain the model to get a one-step SR model with stochastic outputs



Method

北京大学 PEKING UNIVERSITY

Stochasticity Injection | **Advantage Design** | **Policy Optimization**

How to evaluate the advantage of rollout samples?

$$\mathcal{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$$

- Existing metrics: cannot balance fidelity and perception
- Direct combination: cannot adaptively adjust based on image characteristics

Method

北京大学 PEKING UNIVERSITY

Stochasticity Injection | **Advantage Design** | **Policy Optimization**

How to evaluate the advantage of rollout samples?

$$\mathcal{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$$

- Existing metrics: cannot balance fidelity and perception
- Direct combination: cannot adaptively adjust based on image characteristics
- **Dynamic weighting based on metric – texture relations:**
  - FR metrics (PSNR): Smooth regions
  - NR metrics (MANIQA, MUSIQ): Detailed regions

Attribute-aware reward calculation:

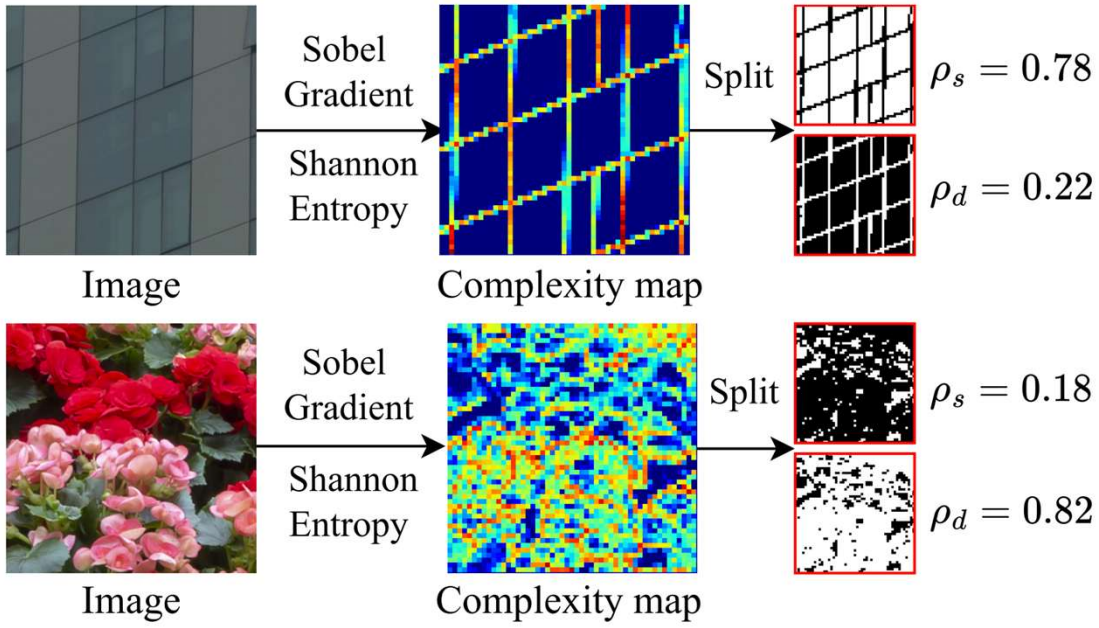
- **Portion of smooth / detailed area**

  as dynamic metric weights

$$R_i = \rho_s \sum_{f \in \mathcal{G}_{FR}} \frac{s_i^f}{|\mathcal{G}_{FR}|} + \rho_d \sum_{f \in \mathcal{G}_{NR}} \frac{s_i^f}{|\mathcal{G}_{NR}|}, i \in [1:G]$$

$$\rho_s = \frac{|\Omega_s|}{(|\Omega_s| + |\Omega_d|)}$$

$$\rho_d = \frac{|\Omega_d|}{(|\Omega_s| + |\Omega_d|)}$$



Method

北京大学 PEKING UNIVERSITY

Stochasticity Injection | Advantage Design | **Policy Optimization**

DPO-style Optimization with Group Rollout:

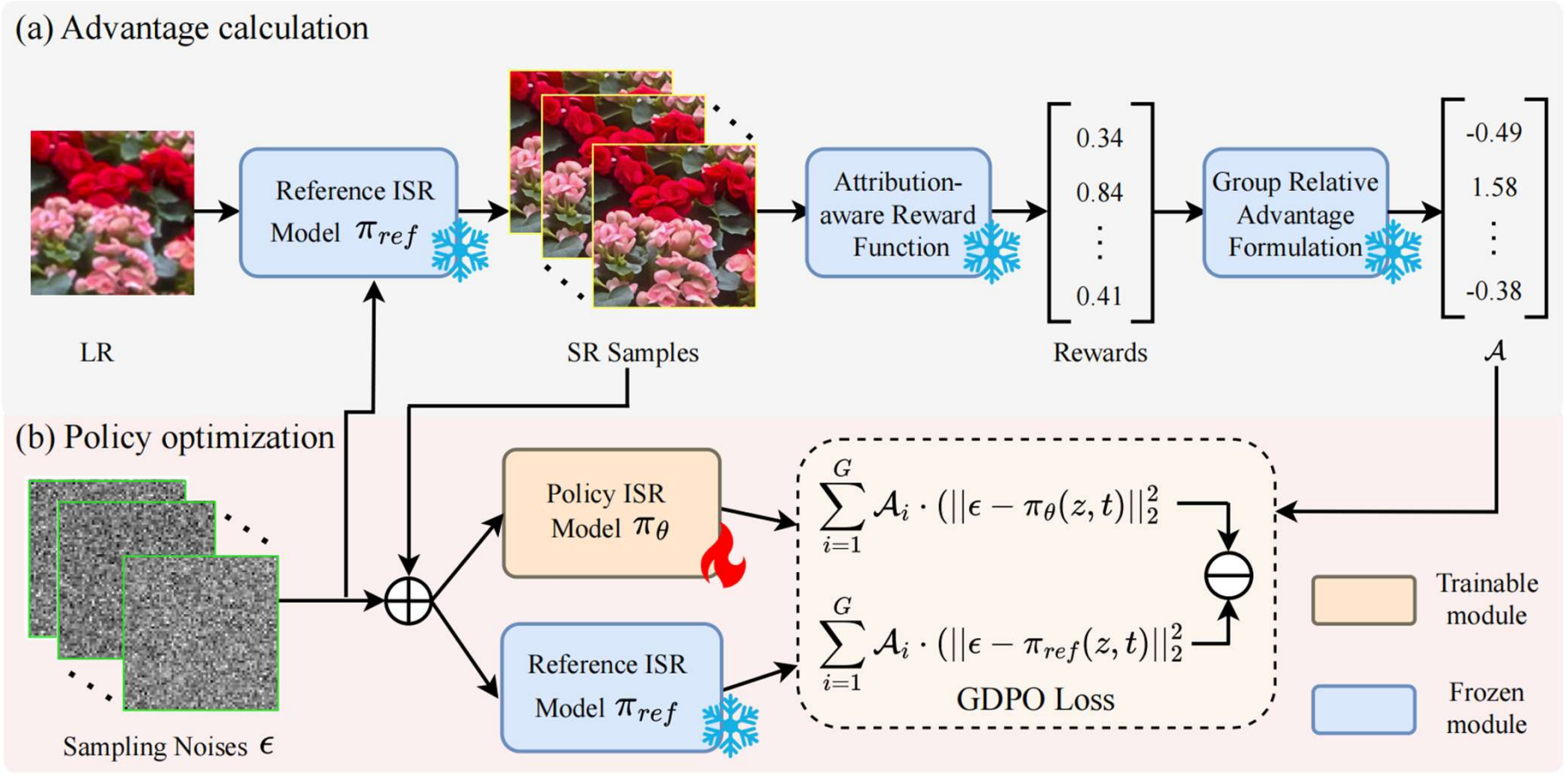
- Log-sigmoid policy gradient from Bradley-Terry model
- Group rollout and relative advantage calculation

$$L(\theta) = -\mathrm{E}_{(x_0^w, x_0^l)\sim\mathcal{D}, x_t^w\sim q(x_t^w|x_0^w), x_t^l\sim q(x_t^l|x_0^l)} \log\sigma(-\omega((\|\epsilon^w - \pi_\theta(x_t^w, t)\|_2^2 - \|\epsilon^w - \pi_{ref}(x_t^w, t)\|_2^2) -(\|\epsilon^l - \pi_\theta(x_t^l, t)\|_2^2 - \|\epsilon^l - \pi_{ref}(x_t^l, t)\|_2^2))),$$

$$\max_{p_\theta} \mathbb{E}_{\{x_{0:T}^i\}_{i=1}^G\sim p_{\theta_{\text{old}}}(\cdot|c)} \left[\sum_{i=1}^{G}\sum_{t=1}^{T} \frac{p_\theta(x_t^i|t,c)}{p_{\theta_{\text{old}}}(x_t^i|t,c)} A_i\right]$$

$$L_{GDPO} = -\mathrm{E}_{x_0\sim\mathcal{D}, x_t\sim q(x_t|x_0)} \log\sigma(-\omega(\sum_{i=1}^{G} \mathcal{A}_i(\|\epsilon - \pi_\theta(x_t, t)\|_2^2 - \|\epsilon - \pi_{ref}(x_t, t)\|_2^2)))$$

Overall Pipeline:



Experiments

北京大学 PEKING UNIVERSITY

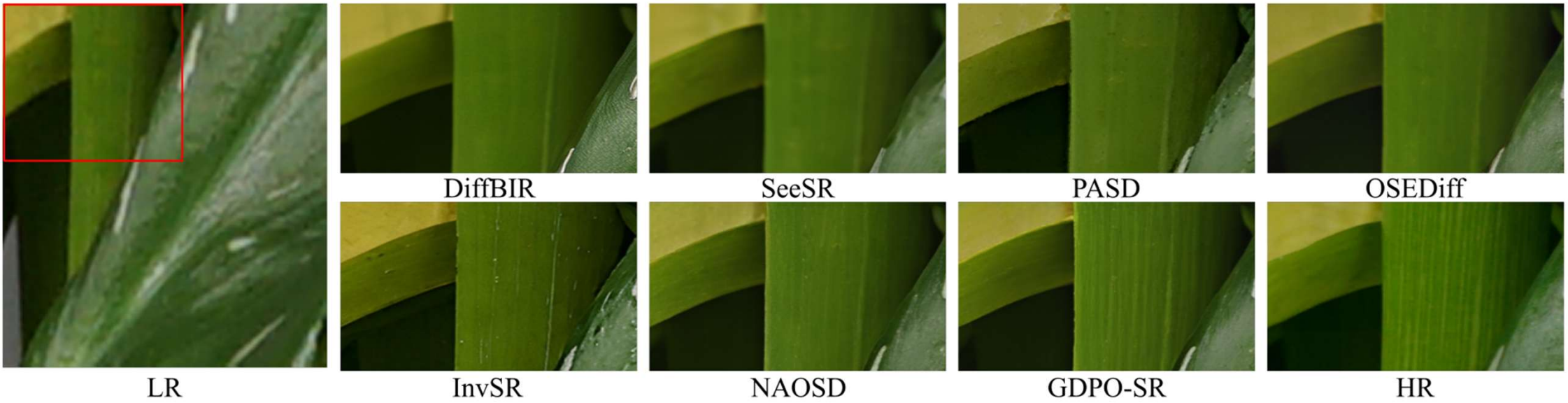
Settings:

- Stable-Diffusion 2.1-based model
- First pretrain the one-step stochastic SR model
- During GDPO-tuning stage, use LoRA to fine-tune both the SD UNet and VAE encoder
- Both pretrain and finetune data are generated from LSDIR & FFHQ with Real-ESRGAN pipeline

Experiments

Qualitative Results：

- Better detail alignment for clear texture



Experiments

北京大学 PEKING UNIVERSITY

Qualitative Results：

- Detail different from HR but more natural when texture is chaotic

Experiments

北京大学 PEKING UNIVERSITY

Quantitative Results：

- Simutanously optimizes FR and NR metrics

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	MANIQA↑	MUSIQ↑	CLIPIQA↑	AFINE↓
DrealSR	StableSR	28.03	0.7536	0.3284	148.98	0.2269	0.5592	58.51	0.6356	35.77
	DiffBIR	26.71	0.6571	0.4557	166.79	0.2748	0.5927	61.07	0.6395	39.42
	SeeSR	**28.07**	0.7684	0.3174	147.39	0.2315	0.6054	**65.08**	0.6905	22.79
	PASD	27.36	0.7073	0.3760	156.13	0.2531	0.6160	64.87	0.6808	33.89
	OSEDiff	27.92	**0.7835**	**0.2968**	**135.30**	**0.2165**	0.5899	64.65	**0.6963**	**21.39**
	InvSR	25.79	0.7176	0.3471	166.42	0.2381	**0.6212**	64.92	0.6918	21.69
	GDPO-SR	**28.18**	**0.7839**	**0.2851**	**138.87**	**0.2112**	**0.6180**	**65.63**	**0.7020**	**18.72**
RealSR	StableSR	24.64	0.7080	0.3002	128.51	0.2140	0.6215	65.88	0.6234	27.62
	DiffBIR	24.75	0.6567	0.3636	128.99	0.2312	0.6252	64.98	0.6463	32.64
	SeeSR	25.15	0.7210	0.3007	125.45	0.2224	0.6441	**69.82**	0.6707	24.07
	PASD	**25.21**	0.6798	0.3380	**124.28**	0.2260	0.6493	68.75	0.6620	39.00
	OSEDiff	25.15	**0.7341**	0.2921	123.49	0.2128	0.6326	69.09	0.6693	20.92
	InvSR	24.30	0.7145	**0.2775**	129.52	**0.2060**	**0.6561**	67.31	**0.6739**	**16.58**
	GDPO-SR	**25.48**	**0.7328**	**0.2675**	**112.13**	**0.1980**	**0.6615**	**69.42**	**0.6760**	**17.73**
DIV2K-val	StableSR	23.26	0.5726	0.3113	**24.44**	0.2048	0.6190	65.92	0.6771	49.31
	DiffBIR	23.64	0.5647	0.3524	30.73	0.2128	0.6209	65.81	0.6704	49.58
	SeeSR	**23.82**	**0.6096**	0.3160	**25.42**	0.1971	0.6187	**68.50**	0.6893	42.18
	PASD	23.23	0.5877	0.3634	30.12	0.2200	**0.6420**	68.41	0.6802	45.19
	OSEDiff	23.72	0.6108	**0.2941**	26.32	**0.1976**	0.6148	67.97	0.6683	44.69
	InvSR	23.10	0.5985	0.3045	28.45	0.1985	0.6406	68.43	**0.7118**	**37.51**
	GDPO-SR	**23.92**	**0.6117**	**0.2897**	26.44	**0.1965**	**0.6423**	**68.80**	**0.6929**	**41.56**

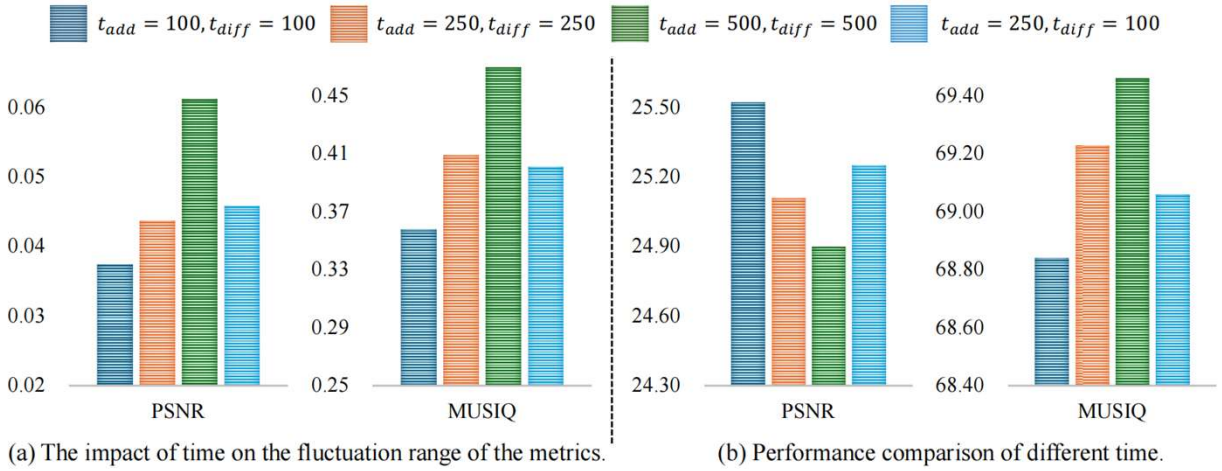
Experiments

Ablation Studies：

- Controlling FR / NR trade-off via noise injection and denoising rate
- t_{add} **encourages exploration**, leading to higher NR and lower FR
- t_{diff} **encourages denoising**, leading to higher FR and lower NR

$$\tilde{z} = \sqrt{\alpha_{t_{add}}} z_{LR} + \sqrt{\beta_{t_{add}}} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

$$z_{SR} = \frac{(\tilde{z} - \sqrt{\beta_{t_{diff}}}\ \mathrm{UNet}(\tilde{z}, c_t, t_{diff}))}{\sqrt{\alpha_{t_{diff}}}}$$



(a) The impact of time on the fluctuation range of the metrics. (b) Performance comparison of different time.

Figure 7. Ablation studies on the timestep setting in NAOSD

北京大学 PEKING UNIVERSITY

Ablation Studies：

- Ablation on GDPO & reward metric components

Table 5. Ablation studies on GDPO on the RealSR dataset.

Method	PSNR↑	FID↓	DISTS↓	MUSIQ↑	AFINE↓
NAOSD	25.25	114.91	0.2001	69.06	20.52
Diffusion-DPO	25.41	112.87	0.2010	69.16	20.22
DanceGRPO	25.10	113.74	0.2049	69.95	16.52
GDPO (ours)	25.48	112.13	0.1980	69.42	18.72

Table 6. Ablation studies on ARF on the RealSR dataset.

Method	LPIPS↓	DISTS↓	MUSIQ↑	CLIPIQA↑
NAOSD	0.2689	0.2001	69.06	0.6617
ARF w/ FR	0.2642	0.1978	67.68	0.6359
ARF w/ NR	0.2866	0.2102	69.80	0.6914
ARF w/o AW	0.2660	0.1979	68.42	0.6331
ARF (ours)	0.2675	0.1980	69.42	0.6760

北京大学 PEKING UNIVERSITY

Ablation Studies：

- Ablation on GDPO & reward metric components

Table 5. Ablation studies on GDPO on the RealSR dataset.

Method	PSNR↑	FID↓	DISTS↓	MUSIQ↑	AFINE↓
NAOSD	25.25	114.91	0.2001	69.06	20.52
Diffusion-DPO	25.41	112.87	0.2010	69.16	20.22
DanceGRPO	25.10	113.74	0.2049	69.95	16.52
GDPO (ours)	25.48	112.13	0.1980	69.42	18.72

Table 6. Ablation studies on ARF on the RealSR dataset.

Method	LPIPS↓	DISTS↓	MUSIQ↑	CLIPIQA↑
NAOSD	0.2689	0.2001	69.06	0.6617
ARF w/ FR	0.2642	0.1978	67.68	0.6359
ARF w/ NR	0.2866	0.2102	69.80	0.6914
ARF w/o AW	0.2660	0.1979	68.42	0.6331
ARF (ours)	0.2675	0.1980	69.42	0.6760

- One-step model w/o GDPO (NAOSD) is already SOTA…

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	MANIQA↑	MUSIQ↑	CLIPIQA↑	AFINE↓
RealSR	StableSR	24.64	0.7080	0.3002	128.51	0.2140	0.6215	65.88	0.6234	27.62
	DiffBIR	24.75	0.6567	0.3636	128.99	0.2312	0.6252	64.98	0.6463	32.64
	SeeSR	25.15	0.7210	0.3007	125.45	0.2224	0.6441	**69.82**	0.6707	24.07
	PASD	**25.21**	0.6798	0.3380	**124.28**	0.2260	0.6493	68.75	0.6620	39.00
	OSEDiff	25.15	**0.7341**	0.2921	123.49	0.2128	0.6326	69.09	0.6693	20.92
	InvSR	24.30	0.7145	**0.2775**	129.52	**0.2060**	**0.6561**	67.31	**0.6739**	**16.58**
	GDPO-SR	**25.48**	**0.7328**	**0.2675**	**112.13**	**0.1980**	**0.6615**	**69.42**	**0.6760**	**17.73**

北京大学 PEKING UNIVERSITY

Quantitative Results：

- Model size & running-time

Table 4. Comparison of mode size, running-time and FLOPs.

Methods	StableSR	DiffBIR	SeeSR	PASD	OSEDiff	InvSR	GDPO-SR
Para.(B)	1.56	1.68	2.51	2.31	1.77	1.33	1.77
Time(s)	10.03	2.72	4.30	2.80	0.11	0.12	0.11
FLOPs(T)	79.94	24.31	65.86	29.13	2.27	2.40	2.27

More Discussions

- Possible false claim in this work: Diffusion-DPO enforces **stronger pixel-level constraint** compare to GRPO

ment biased toward the higher-reward side. In contrast to GRPO, which requires computing the full-image likelihood, GDPO inherits Diffusion-DPO's advantage of implicit likelihood computation while imposing pixel-level constraints, thereby learning local details more effectively.

$$L(\theta) = -\mathrm{E}_{(x_0^w,x_0^l)\sim\mathcal{D},x_t^w\sim q(x_t^w|x_0^w),x_t^l\sim q(x_t^l|x_0^l)} \log\sigma(-\omega((\|\epsilon^w - \pi_\theta(x_t^w,t)\|_2^2 - \|\epsilon^w - \pi_{ref}(x_t^w,t)\|_2^2) -(\|\epsilon^l - \pi_\theta(x_t^l,t)\|_2^2 - \|\epsilon^l - \pi_{ref}(x_t^l,t)\|_2^2))),$$

$$\mathcal{L}_{\text{GRPO}}^{\text{diff}} = -\mathbb{E}_{\{x_{0:T}^i\}_{i=1}^G\sim p_{\theta_{\text{old}}}(\cdot|c)} \left[\sum_{i=1}^G\sum_{t=1}^T A_i \frac{p_\theta\left(x_t^i \mid t,c\right)}{p_{\theta_{\text{old}}}\left(x_t^i \mid t,c\right)}\right]$$

$$\sim -\mathbb{E}_{\{x_{0:T}^i\}_{i=1}^G\sim p_{\theta_{\text{old}}}(\cdot|c)} \left[\sum_{i=1}^G\sum_{t=1}^T A_i \exp(\|\epsilon_t^i - \pi_\theta(x_t^i,t)\| - \|\epsilon_t^i - \pi_{\text{ref}}(x_t^i,t)\|)\right]$$

- Both Diffusion-DPO and GRPO for Diffusion calculates likelihood in the form of pixel-level L2 distance

Take-aways

- **Leverage stochasticity to optimize SR preference via RL**
  - Add noise to one-step model for stochasticity injection
  - Reward weighting via texture region portion
  - Aggregate DPO and GRPO policy optimization
- **Simutanously optimizes fidelity & perception**
  - Simutanously optimizes FR & NR
  - Retain natural detail when content is hard to restore

Thanks for listening!

Presenter: Jinyi Luo
2026.03.29