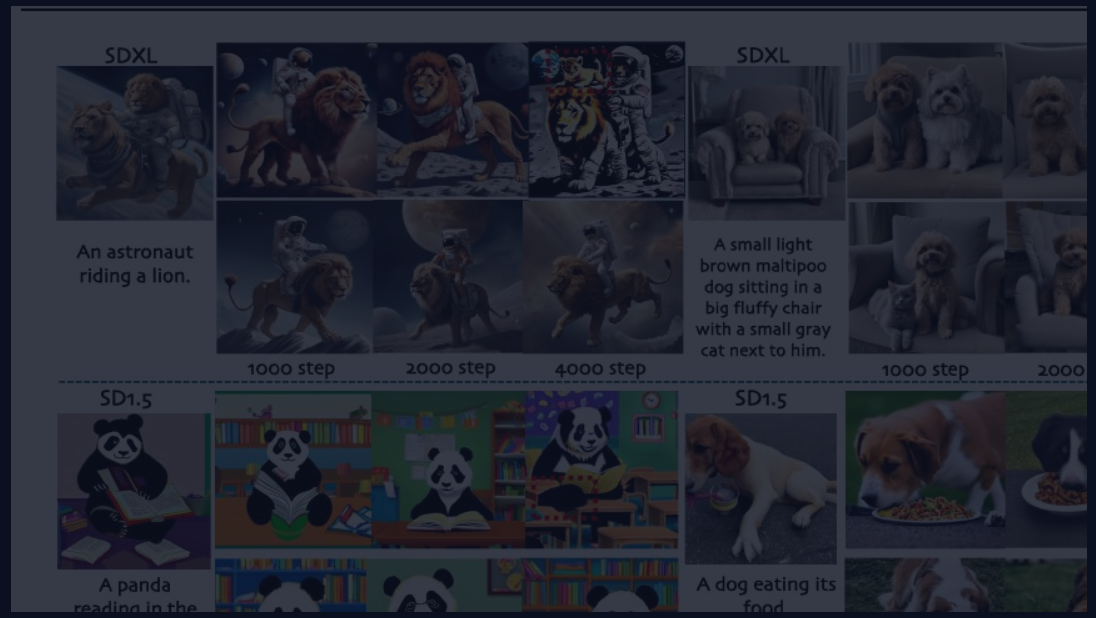
SIPO

Stabilized and Improved Preference Optimization

for Aligning Diffusion Models

Paper Presentation Deck **Image + Video Preference Alignment**

Key idea: use timestep–aware importance reweighting to stabilize diffusion preference optimization and reduce off–policy bias from offline preference data.



Xiaomeng Yang et al. · ICML 2026

One-slide summary: What problem does this paper solve?

SIPO targets two root causes: training instability and offline distribution shift.

Problem

Diffusion-DPO on diffusion models often shows loss rebound, late-stage performance drops, and high sensitivity to β.

Finding

Early timesteps have low importance weights and produce noisy gradients; middle-to-late timesteps are more useful.

Method

DPO-C&M performs timestep-aware clipping/masking; SIPO adds clipped importance reweighting.

Results

More stable on SD1.5/SDXL images and CogVideoX/WanX videos, with overall gains over Diffusion-DPO.

Take-home message

Preference optimization for diffusion models must consider not only sample preference, but also whether each denoising timestep is reliable.

Background: Why do diffusion models need preference optimization?

The goal is to make image/video generation better match human preferences, not just fit the data distribution.

DPO is natural for language models, but diffusion models are harder.

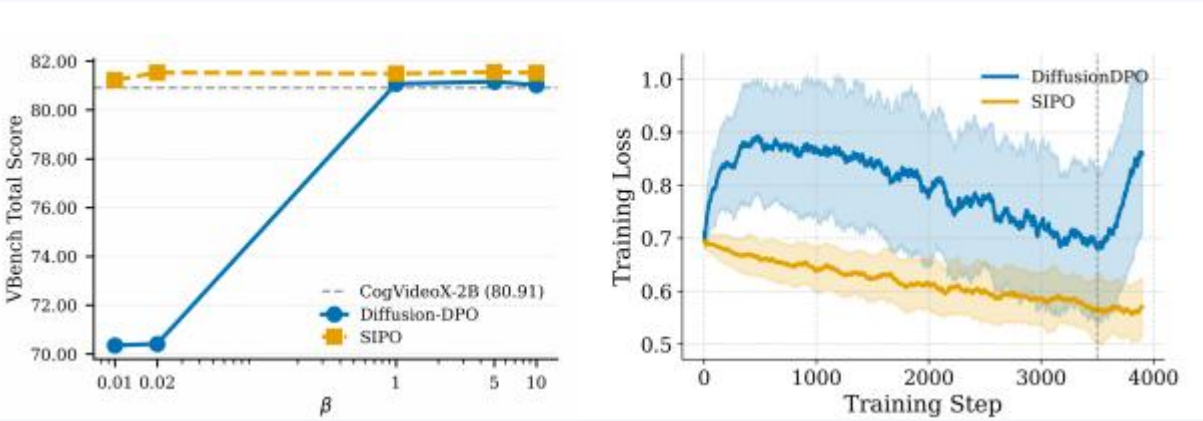
Diffusion generation follows a long denoising trajectory: the same preference pair is mapped to different noise steps, and gradient quality changes strongly with timestep.

Offline preference data is not sampled from the current policy.

When data comes from a fixed or stronger model, the training distribution q differs from the current model pθ, causing off–policy drift.

Challenge 1: Diffusion–DPO training is unstable.

The paper attributes instability to β sensitivity, loss rebound, and late–stage test accuracy decline.



β sensitivity

Too small: overly aggressive updates and performance drops; too large: overly conservative updates and limited gains.

Non–monotonic loss

Diffusion–DPO can show late–stage training loss rebound and even collapse.

Late–stage evaluation drop

Test accuracy curves show that longer training is not always better; reward hacking or overfitting may occur.

Key issue

DPO aims to increase positive rewards, decrease negative rewards, and widen the gap, but vanilla Diffusion–DPO often fails to follow this reward dynamic.

Key observation: Not all timesteps deserve equal training weight.

Early timesteps have low importance weights and contribute more noisy gradients.

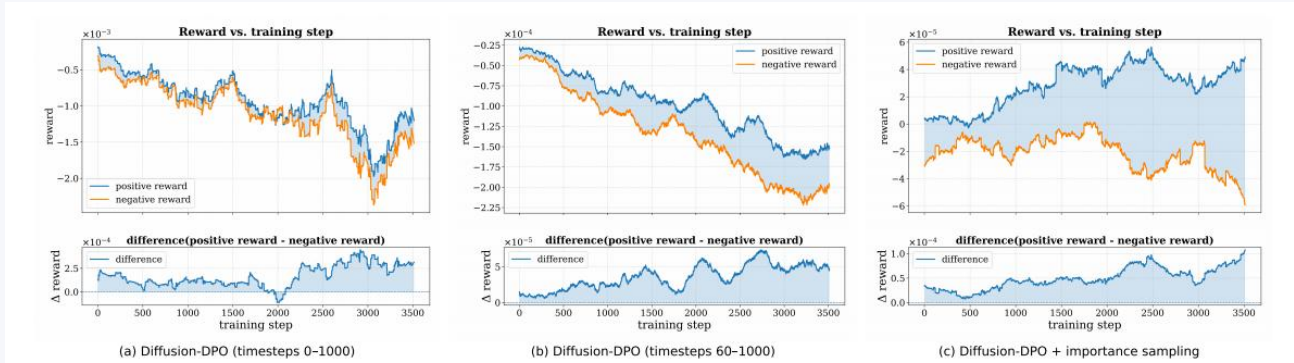
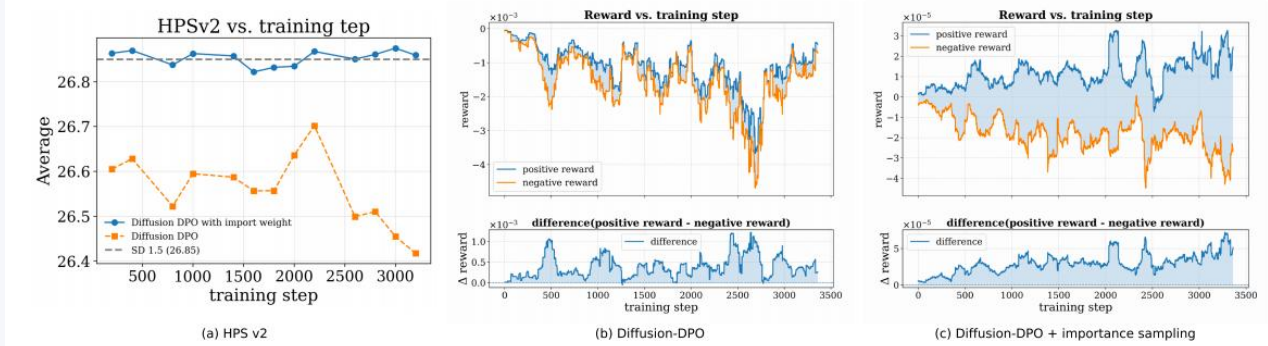


Figure 2. Reward dynamics in Diffusion-DPO under different training settings. (a) Random timesteps (0–1000) give unstable rewards and a small gap. (b) Restricting to 60–1000 stabilizes training but reduces rewards, while enlarging the gap. (c) Dynamic timestep clipping with importance sampling further enlarges the gap, with positive rewards increasing and negative rewards decreasing, better matching the DPO objective.



Empirical finding

Removing low–weight timesteps enlarges the positive—negative reward gap and better matches the DPO objective.

Timestep interpretation

$w(t)<0.9$ mainly appears in $t\in[0,63]$; middle–to–late timesteps are more stable and informative.

Method insight

Preference optimization should use timestep–aware filtering/reweighting instead of uniform timestep sampling.

Challenge 2: Off–policy mismatch from offline preference data.

The training distribution q differs from the current model distribution pθ, and pθ keeps drifting during training.

Preference data q(x|c)

Often comes from a fixed reference or stronger models, such as Flux / DALL·E 3 generated data.

Current model pθ(x|c)

Keeps changing during training; optimizing only on offline data accumulates distribution drift.

Observed behavior

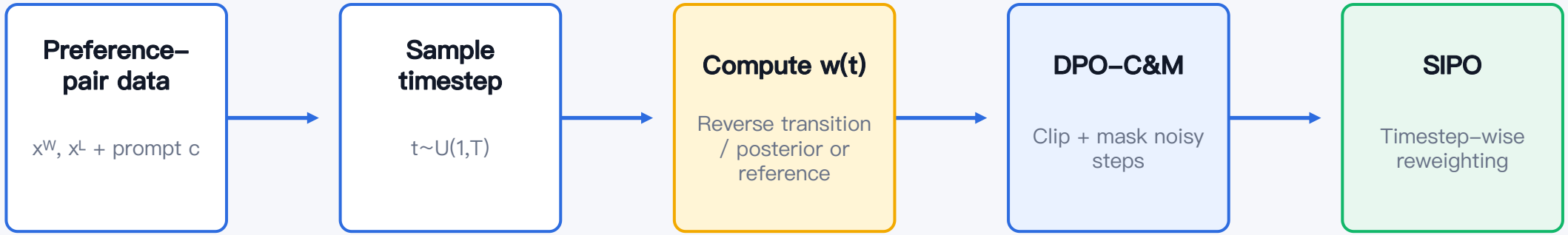
HPSv2 drops; reward curves oscillate and the positive—negative gap remains small.

Role of importance sampling

Use likelihood ratios to reweight offline samples, making updates closer to reliable regions of the current policy.

Method overview: SIPO = C&M for stability + importance reweighting for bias correction.

Two–stage logic: filter low–quality timesteps first, then adaptively reweight reliable updates.



What does DPO–C&M solve?

Suppresses early low–weight timesteps and noisy gradients, preventing Diffusion–DPO from being driven by unreliable updates.

What does SIPO further solve?

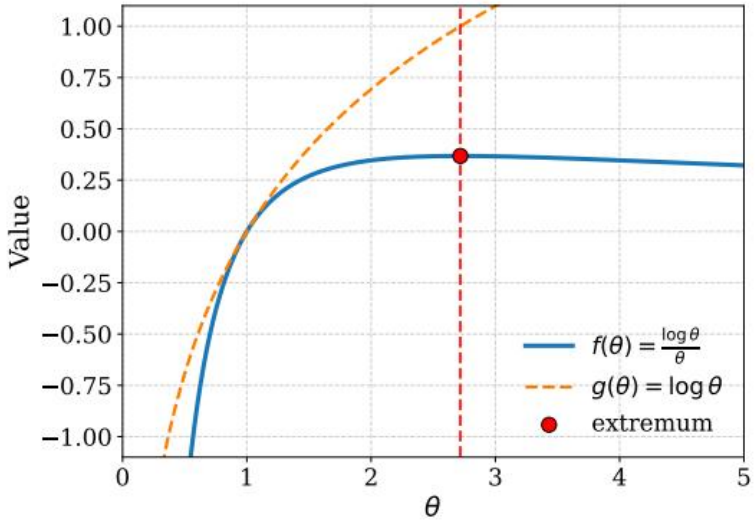
Corrects off–policy bias via clipped importance reweighting, reducing β sensitivity and over–optimization.

DPO–C&M: Use clipped importance weights as soft filtering.

w(t) serves both as gradient scaling and a soft mask.

$$\tilde{w}(t) = \mathrm{clip}(w(t), 1 - \varepsilon, 1 + \varepsilon)$$

$$L_{\text{DPO–C\&M}} = -\,\mathbb{E}\left[\, \tilde{w}(t) \cdot \log \sigma\left(-\beta T \omega(\lambda_t) \Delta \ell(\theta) \right) \right]$$



1. Low–weight timesteps: gradients are down–weighted or masked.
2. Reliable middle–to–late timesteps: preserve the main preference signal.
3. Clipping prevents extreme weights from amplifying variance.

Intuition

Do not force the model to learn preferences at low–confidence timesteps; this step mainly improves stability.

SIPO: Timestep-wise clipped importance reweighting.

Rewrite offline preference learning as an importance-weighted objective to reduce off-policy bias.

$$\max \; \mathbb{E}\left[w_\theta \cdot r(c, x_0) \right] - \beta \, \mathrm{KL}\left[p_\theta(x_0|c) \,\|\, \mathrm{pref}(x_0|c) \right]$$

$$L_{\mathrm{SIPO}}(\theta) = -\mathbb{E}\left[\log \sigma\left(w_\theta(t) \cdot (\psi^{w}_t - \psi^{l}_t) \right) \right]$$

1. Correct distribution shift

Use the pθ/q idea to reweight offline preference data toward more reliable policy update directions.

2. Control update magnitude

Clip weights to prevent certain timesteps or samples from producing excessively large gradients.

3. Reduce β sensitivity

The weighted log-ratio has a soft upper bound, mitigating over-optimization and reward hacking.

Core story

C&M decides which timesteps should not be learned; SIPO decides how much each timestep should learn under offline data.

Experimental setup: Images + videos + AI feedback.

Covers SD1.5/SDXL, CogVideoX, and WanX, validated by automatic metrics and human evaluation.

Task	Model	Data / preference source	Metrics
Text–to–Video	CogVideoX–2B/5B Wan2.1–1.3B	10k high–quality video preference pairs	VBench Total / Quality / Semantic
Text–to–Image	SD1.5 / SDXL	Pick–a–Pic pairwise preferences	PickScore, HPSv2, ImageReward, Aesthetics
AI Feedback	FLUX–dev	YOLO–World generates 5k preference pairs	GenEval multi–task accuracy
Training	16×A100	Adam, lr=1e−5, mixed precision	Diffusion–DPO uses its best β; SIPO uses stable β settings.

Fairness
Diffusion–DPO is reported with its best β from sensitivity analysis; SIPO remains more robust.

Generality
SIPO is validated on diffusion and flow–matching image models, as well as video models.

Stability
Beyond final metrics, the paper tracks training dynamics, hyperparameter sensitivity, and human ranking.

Video generation results: SIPO leads stably on VBench.

Especially under long training, Diffusion–DPO may collapse, while SIPO remains stable.

Model	Pretrained	Diffusion–DPO	DPO–C&M	SIPO
CogVideoX–2B @500	80.91	81.16	81.37	**81.53**
CogVideoX–2B @1000	80.91	67.28	81.46	**81.68**
CogVideoX–5B	81.91	82.02	82.17	**82.28**
WanX–1.3B	84.26	84.41	84.54	**84.78**

67.28

Diffusion–DPO @1000 steps shows severe collapse

81.68

SIPO under the same setting keeps stable improvement

Conclusion

C&M already stabilizes training; SIPO improves further, showing that reweighting is not only filtering but also improves alignment.

Image generation results: SIPO is the most balanced overall.

Best on PickScore, HPSv2, and Aesthetics, and better than DPO–C&M on all metrics.

Method	PickScore	HPSv2	ImageReward	Aesthetics
SD V1.5	20.73	0.2341	0.1697	5.337
Diffusion–DPO	20.97	0.2656	0.2989	5.594
Diffusion–KTO	21.15	0.2719	0.6156	5.697
DPO–C&M	21.57	0.2744	0.3024	5.783
SIPO	**21.68**	**0.2783**	**0.3051**	**5.811**

vs. Diffusion–DPO

PickScore 20.97 → 21.68
HPSv2 0.2656 → 0.2783
Aesthetics 5.594 → 5.811

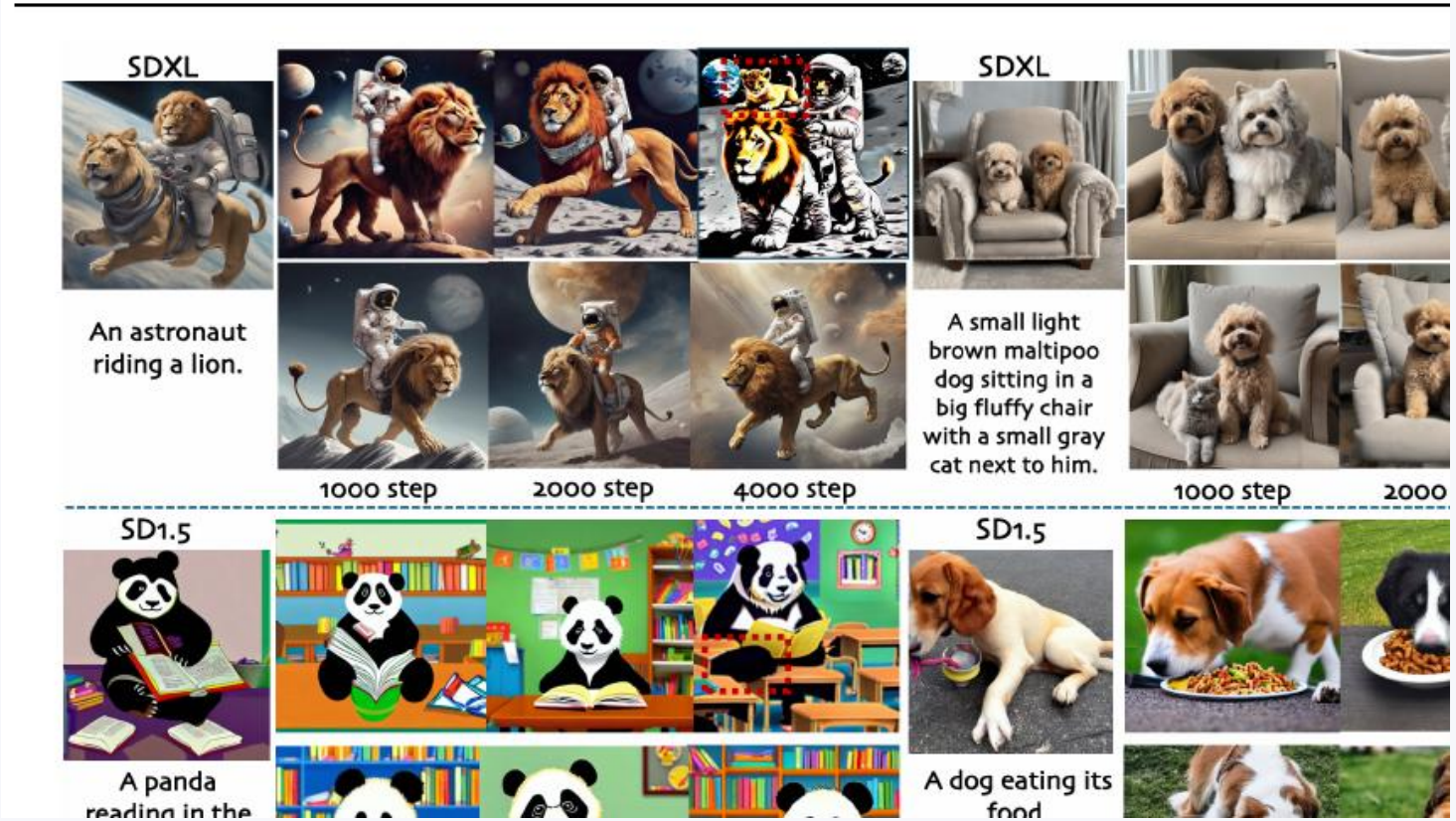
vs. DPO–C&M

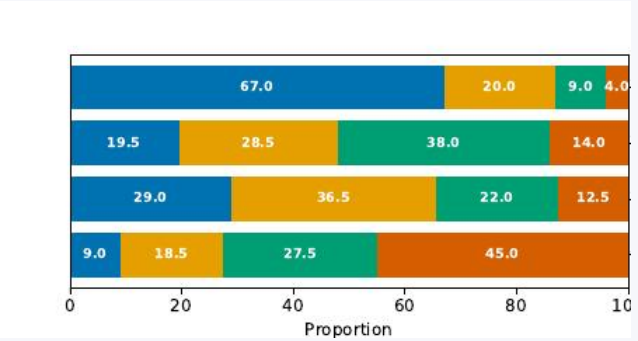
SIPO further improves all four metrics, indicating extra gains from timestep–wise reweighting.

Note: Diffusion–KTO has higher ImageReward, but SIPO is more balanced across preference and aesthetic metrics.

Qualitative results and human evaluation: SIPO better preserves semantics and visual quality.

Visual comparisons and human ranking both support the stability conclusion.





Visual observations

Under long training, Diffusion–DPO may show consistency, color, and artifact issues; SIPO remains more stable at 1000/2000/4000 steps.

Human evaluation

Across 200 prompts, SIPO achieves the highest Rank–1 share: 67%.

Conclusion: Timestep–aware alignment is key to diffusion preference optimization.

Contribution 1

Systematically analyzes Diffusion–DPO instability and identifies low–weight, high–noise early timesteps.

Contribution 2

Proposes DPO–C&M to stabilize training via timestep–aware clipping and masking.

Contribution 3

Proposes SIPO to mitigate off–policy bias with clipped importance reweighting and improve robustness.

AI feedback results

On GenEval, SIPO reaches Overall=0.85, outperforming FLUX.1–dev / SFT / D3PO / Diffusion–DPO.

Final takeaway

Preference optimization should consider both the sample preference signal and denoising–timestep reliability. SIPO connects them with importance weights.