NEURIPS 2026 · POSTER

Consistency-enforced Preference-driven Uunlearning (CePU)

Enabling Preference-driven Unlearning in Few-step Distilled Text-to-Image Diffusion Models

Gaurav Patel · Jun Fang · Greg Ver Steeg · Qiang Qiu · Sravan Sripada
Purdue University · Amazon AGI

THE PROBLEM

Few-step models are what ships, and they leak more

Different dynamics

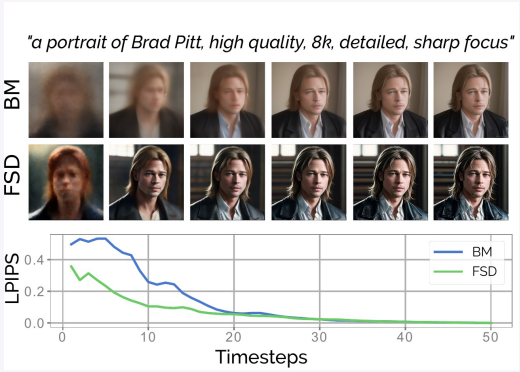
Distillation cuts 50+ steps to 2–8. FSD models commit to an image within the first few steps.

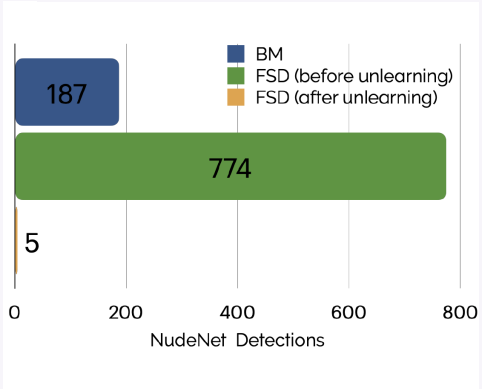
Amplified risk

On I2P prompts, NudeNet detections rise from 187 (base) to 774 (FSD).

Re-distilling is costly

Unlearn the base model, then distill again: tens to thousands of A100 GPU-hours, with no guarantee the edit survives.





KEY OBSERVATION

Noise-prediction rewards don't fit few-step models

DPO-style unlearning uses ε-error

Diffusion-DPO, DUO and SafetyDPO rank preferred vs. unwanted samples by noise-prediction error.

The proxy breaks for FSD models

ϵ-error $\propto x_0$-error only when $c_{\text{skip}} = 0$ (Lemma 1). Consistency-style models (LCM) use $f_\theta^{(t)}(x_t) = c_{\text{skip}} \cdot x_t + c_{\text{out}} \cdot F_\theta^{(t)}(x_t)$, so the link is lost.

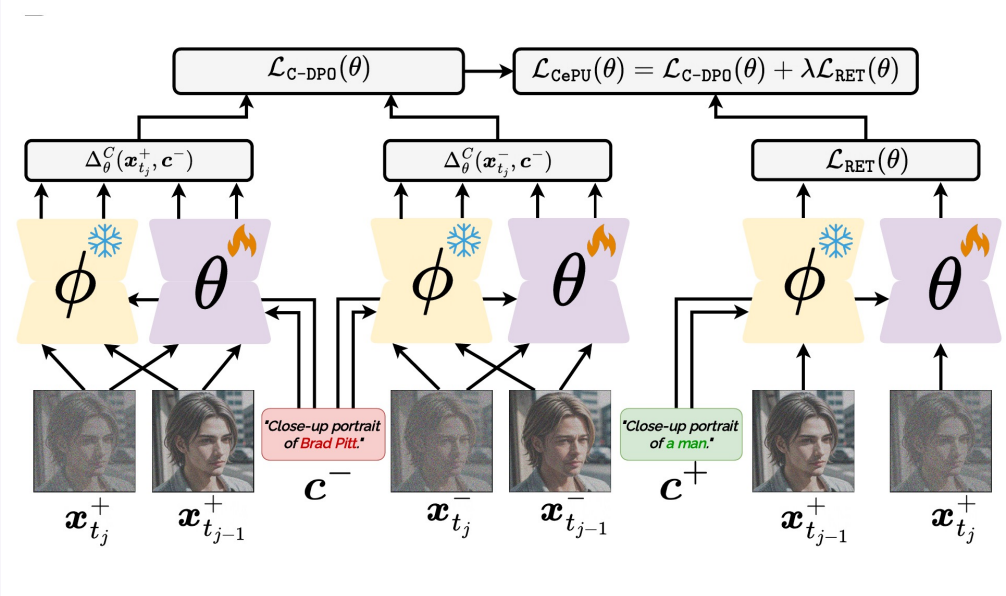
Plain x_0-error is not enough

It ignores the few-step inductive bias. We call this mismatch **objective dissonance**.



METHOD

CePU rewards consistency instead of noise error



Consistency delta

$$\Delta_\theta^C = \left\| f_\theta^{t_j}\left(x_{t_j}\right) - f_\theta^{t_{j-1}}\left(x_{t_{j-1}}\right) \right\|$$

– (same for frozen ϕ)

Preference loss

Lower Δ_θ^C on the safe sample x^+ than on the unwanted x^-, both under the unwanted prompt c^-.

Final objective

$$\mathcal{L}_{\text{CePU}}(\theta) = \mathcal{L}_{\text{C-DPO}}(\theta) + \mathcal{L}_{\text{RET}}(\theta)$$

$\mathcal{L}_{\text{RET}}(\theta)$ anchors safe-prompt outputs to the frozen model.

RESULTS · IDENTITY UNLEARNING

Removes the target identity, keeps everyone else

Harmonic mean ↑	SafetyDPO	CePU
V8 · Jolie	49.4	71.0
V8 · Pitt	63.5	73.9
V7 · Jolie	59.4	61.6
V7 · Pitt	66.2	70.5

Harmonic mean of forgetting (1 – target accuracy) and accuracy on other identities, DreamShaper-V8/V7-LCM. CePU is best among preference-based methods in all four settings, with the lowest LPIPS in three.



RESULTS · NUDITY UNLEARNING

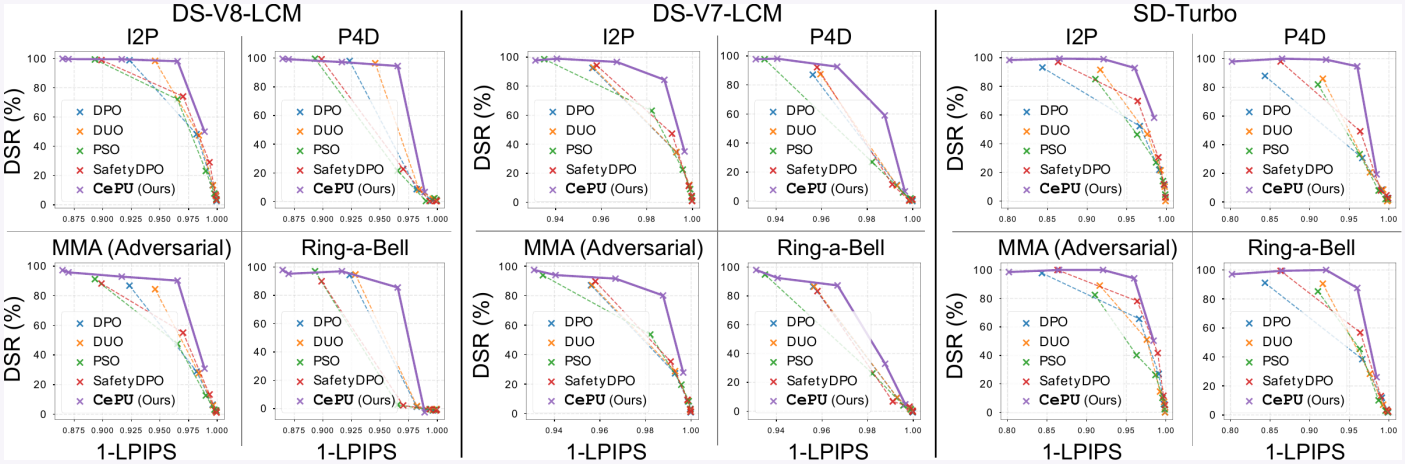
Robust to unseen red-teaming prompts

876 → 6
NudeNet detections on SD-Turbo, across six red-teaming sets

99.3%
Defense success rate, SD-Turbo (best overall)

95.8%
Defense success rate, DS-V8-LCM (best preference-based)

0.079
LPIPS on SD-Turbo, lowest among preference-based



PRACTICALITY

Cheap, direct, and general

~15 min

on one A5000 GPU

vs. ~24 h on 4×A5000 to unlearn the base model and re-distill.

19 vs 20

I2P detections

Unlearning the FSD model directly matches unlearn-then-distill.

75.7%

Object unlearning

Imagenette harmonic mean vs. 74.4% for FADE, with lower LPIPS (0.09 vs. 0.12). No preference pairs needed.

1 step

Still effective

CePU keeps a strong safety–utility trade-off even with single-step sampling.

TAKE-HOME

Unlearn the model you deploy. For few-step models, reward consistency, not noise-prediction error.

CePU removes identities, nudity and objects directly from distilled models in minutes, with a better safety–utility trade-off than existing preference-based methods.