NeurIPS 2026

Poster

D-OPSD: On-Policy Self-Distillation for Continuously Tuning Step-Distilled Diffusion Models

Dengyang Jiang 1,2 Xin Jin 2 Dongyang Liu 4,2 Zanyi Wang 3 Mingzhe Zheng 1 Ruoyi Du 2

Xiangpeng Yang 2 Qilong Wu 2 Zhen Li 2 Peng Gao 2* Harry Yang 1* Steven C. H. Hoi 2

*1 HKUST 2 Alibaba Group 3 UCSD 4 CUHK * Corresponding authors*

Project: https://vvvvvjdy.github.io/d-opsd

RedNote

Background: The Challenge of Continual Tuning

Step-distilled diffusion models generate high-quality images in only a few steps, but conventional supervised fine-tuning (SFT) trains on noised target images rather than the model's own sampling states.

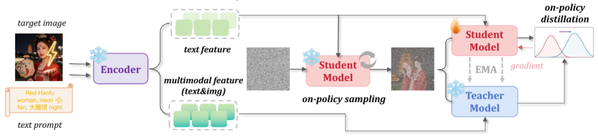
This mismatch can disrupt the distilled sampling dynamics and degrade few-step generation quality. Online RL follows model roll-outs, but typically requires a reward function.

Observation: In-Context Capability of Diffusion Models



With an LLM/VLM encoder, adding a target image to the prompt condition can already guide the frozen model toward the target concept or style.

Method: On-Policy Self-Distillation



Student: samples a few-step trajectory using text-only conditioning. **Teacher:** uses the same states with a richer text + target-image condition.

The student matches EMA teacher velocities on its own roll-outs. After training, inference uses the original few-step text-to-image pipeline.

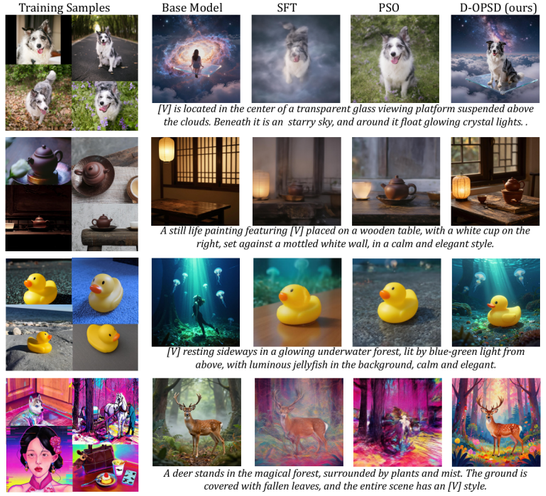
Results: Full Fine-Tuning on Two Few-Step Models

Table 2: **System-level comparison against baseline methods in full-finetuning settings.** The **best** and second-best results on each metric are highlighted in bold and underlined.

Method	FID↓	DINO-D↓	LPIPS-D↓	Quality-S↑	Aesthetic-S↑	GenEval↑	DPG↑
Z-Image-Turbo							
Base-Model	48.6858	0.1274	0.7036	3.7624	3.5218	**0.7543**	**84.7645**
Vanilla SFT	82.2036	0.1896	0.6787	2.6121	2.4852	0.1588	69.9746
PSO	88.4343	0.1716	0.6530	2.8653	2.6424	0.2475	72.7363
D-OPSD (ours)	**40.4938**	**0.1088**	**0.6419**	**3.8438**	**3.6195**	0.7170	84.1116
FLUX.2-klein							
Base Model	45.2335	0.1210	0.6894	3.7626	3.4465	**0.8155**	**85.5624**
Vanilla SFT	98.7893	0.2565	0.6567	2.5038	2.4121	0.1113	56.2138
PSO	82.1438	0.2185	0.6511	2.8066	2.5952	0.2406	63.4174
D-OPSD (ours)	**38.2932**	**0.0902**	**0.6476**	**3.8328**	**3.5511**	0.7298	83.8927

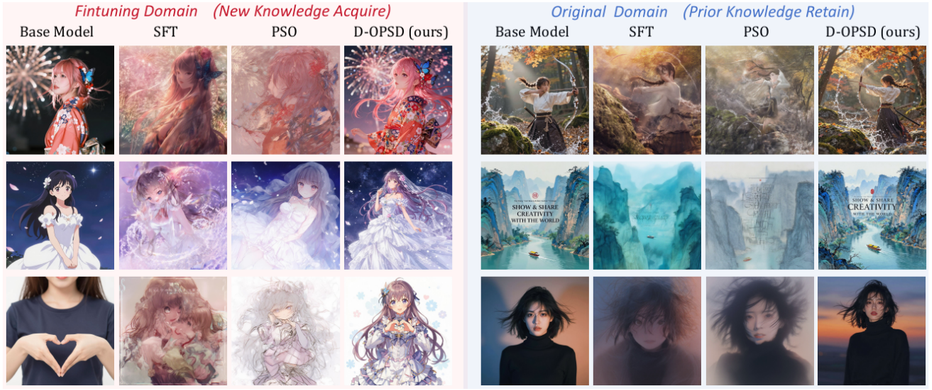
D-OPSD adapts to the target domain while preserving generation quality and much of the original knowledge. On Z-Image-Turbo, GenEval is **0.717**, versus **0.159** for SFT and **0.754** for the base model.

LoRA: Learning New Concepts and Styles



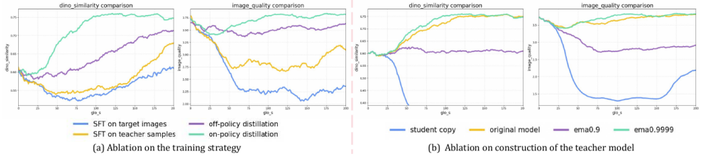
D-OPSD learns the target identity or style and generalizes to new prompts while maintaining few-step image quality.

Full Fine-Tuning: New Domain and Prior Knowledge



Compared with SFT and PSO, D-OPSD adapts toward the anime domain while retaining stronger performance in the original domain.

Ablation and Takeaway



On-policy distillation achieves faster target alignment and maintains generation quality. An EMA teacher with momentum 0.9999 provides the strongest result among the teacher variants studied.

Takeaway: Paired image-text data can supervise continual tuning on the student's own trajectory, without a separate reward model or an extra training stage.

Limitation: The method depends on the base model's ability to generate useful supervision under multimodal conditioning.