Qualcomm

Cross-Space Distillation

Teaching One-Step Students with Modern Diffusion Teachers

Anh Nguyen[1*] **Ngan Nguyen**[1*] **Duc Vu**[1*] Trung Dao[2] Viet Nguyen[3] Quan Dao[4] Kien Nguyen[1]
Chi Tran[1] Phong Nguyen[1] Khoi Nguyen[1] Cuong Pham[1] Dimitris N. Metaxas[4] Vishal M. Patel[3] Anh Tran[1]

[1] Qualcomm AI Research [2] University of Wisconsin–Madison [3] Johns Hopkins University [4] Rutgers University

ECCV 2026

 * Denotes equal contribution arXiv:2606.32020

Motivation: one-step models are what we deploy

The backbones small enough to ship are generations behind the models that set quality.

WHAT WE DEPLOY

SD 1.5, SD 2.1

512 × 512

0.86 B parameters

1 step

distil

WHAT SETS QUALITY

SD 3.5, FLUX.2, Kolors

1024 × 1024

up to 4 B parameters

20 – 50 steps

Can a deployable Student learn from these Teachers?

Mismatched latent spaces cause existing distillation methods to fail

Distribution matching (DMD2, SiD) needs both models in one latent space.

64 × 64 × 4

Student — SD 1.5, 512 × 512

128 × 128 × 16

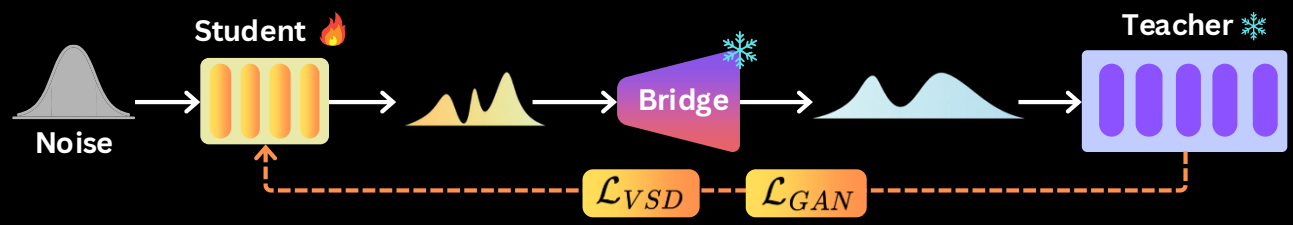
Teacher — SD 3.5, 1024 × 1024

We call this setting **Cross-Space Distillation**

Latent resolution and VAE parameterization differ at the same time.

Idea: a Bridge between the two spaces

One stage per axis.



Where the Bridge sits in the distillation loop.

the Bridge

resolution mismatch

VAE mismatch

Student latent → **spatial alignment** → **semantic alignment** → **Teacher latent**

free · no new parameters

learned · ≈ 5 M parameters

Training the Bridge

Two terms, both computed in Teacher space.

Match the Teacher's latent

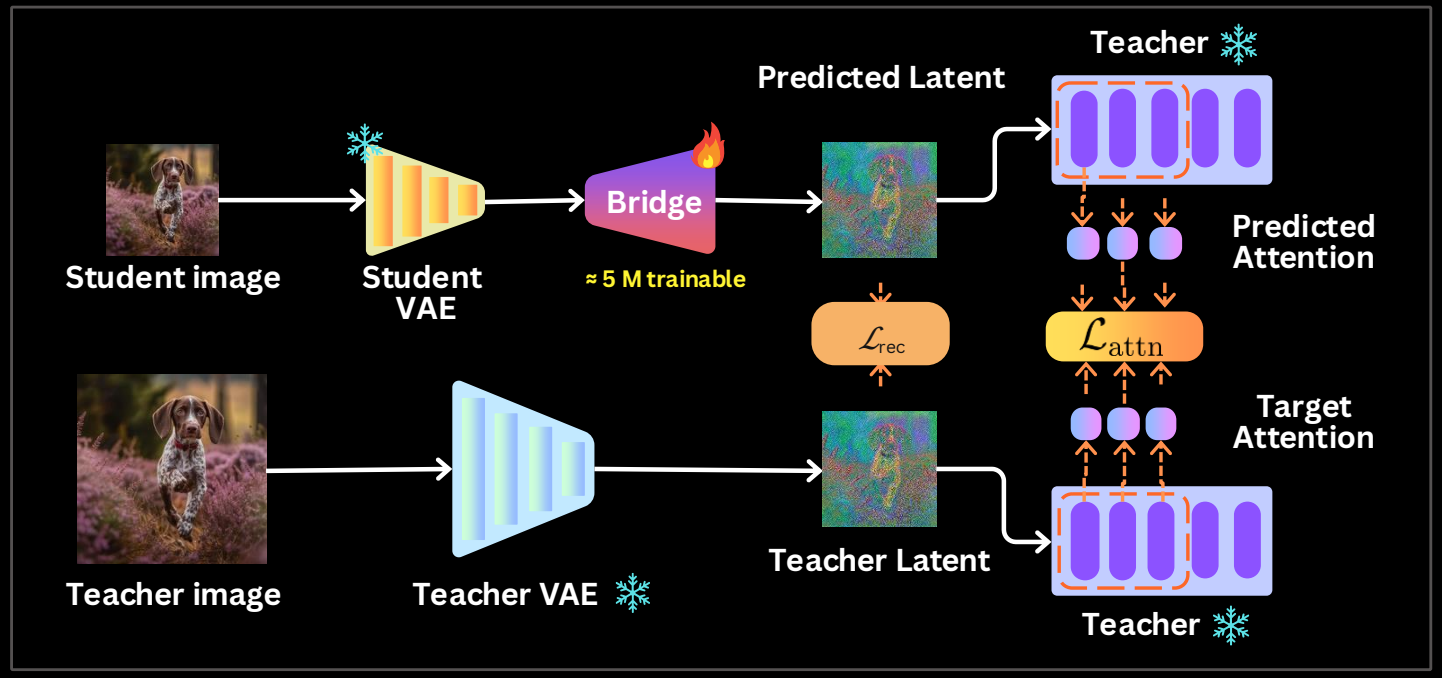
L1 distance between the mapped Student latent and the true one

Match the Teacher's attention

reverse KL between the Teacher's self-attention on each latent

HOW THE BRIDGE IS TRAINED

≈ 2 M images · 8 × H100, 20 h



Both terms are measured in the Teacher's space.

Ablation: the choice of training objective

Same Student, same Teacher, same Bridge. Only the objective changes.

	L1 ↓	PSNR ↑	SSIM ↑
Reconstruction only per-position signal only	0.041	23.58	0.70
+ E-LatentLPIPS no measurable gain	0.042	23.62	0.71
+ attention fidelity (ours) adds the long-range signal	0.035	24.97	0.75

Measured after the Teacher's decoder. L1 reconstruction in every row.

WHY BOTH TERMS

+1.39 dB

PSNR, over reconstruction alone

Reconstruction is a per-position signal. Attention fidelity is what supplies the long-range one.

E-LatentLPIPS: + 0.04 dB.

Result: the same Student, five modern Teachers

One SD 1.5 backbone, one step. Only the Teacher changes.

baseline

taught by a modern Teacher

SD 1.5 DMD2

5.37

+ SDXL

9.04

+ Kolors

9.33

+ PixArt-Σ

8.65

+ SD 3.5 Medium

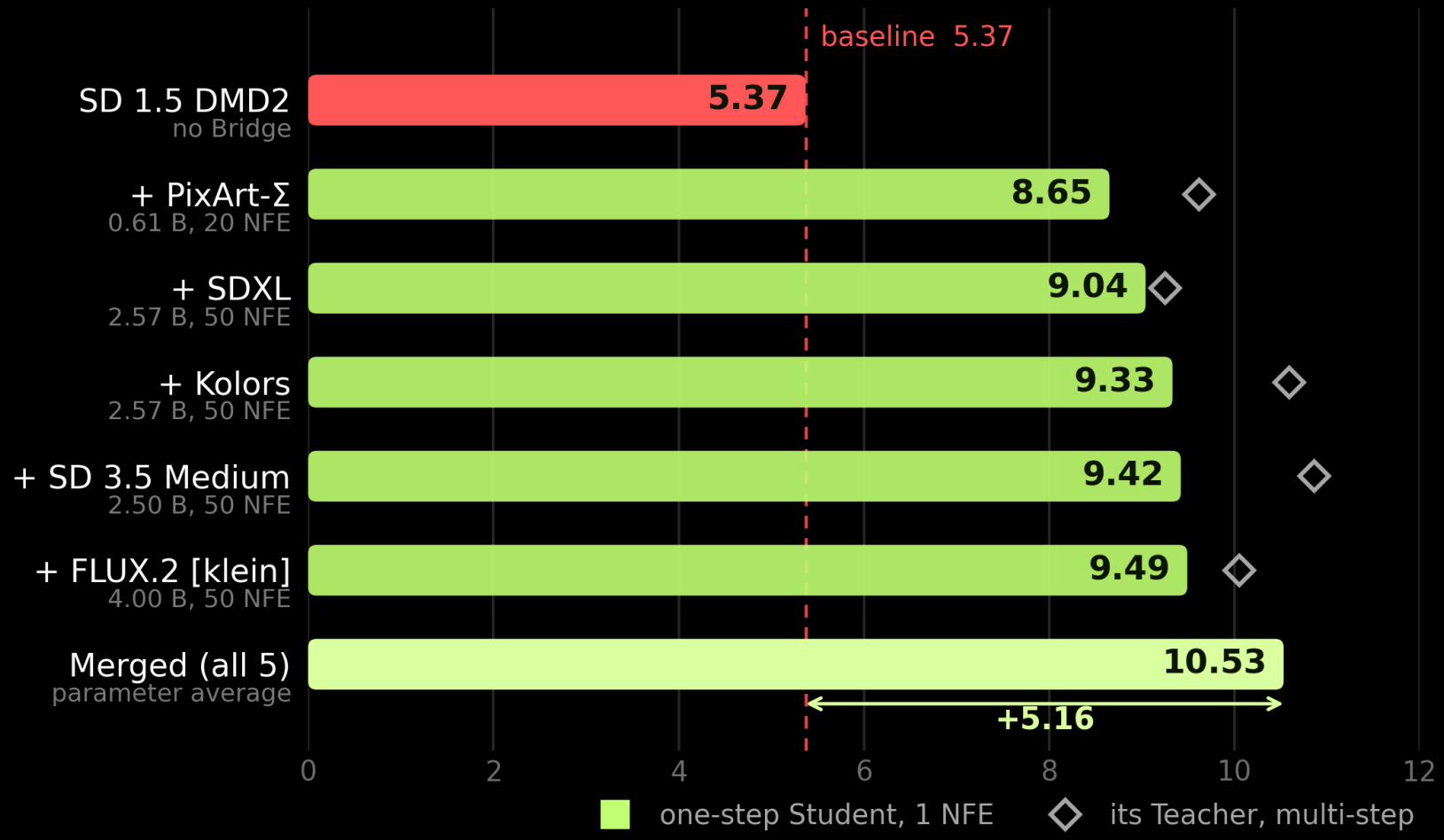
9.42

+ FLUX.2 [klein]

9.49

One prompt. Scores are HPSv3 over the benchmark's 12,000 prompts.

Result: HPSv3 across all five Teachers



Every Teacher gives at least **+3.28** over the baseline.

10.53

merged Student, one step

Above SDXL (2.57 B) and FLUX.2 [klein] (4.00 B) at fifty steps.

Model merging

Average the parameters of the five distilled Students.

SDXL | Kolors | PixArt-Σ | SD 3.5 M | FLUX.2 → **one Student**

average

One-step samples from the merged SD 1.5 Student.

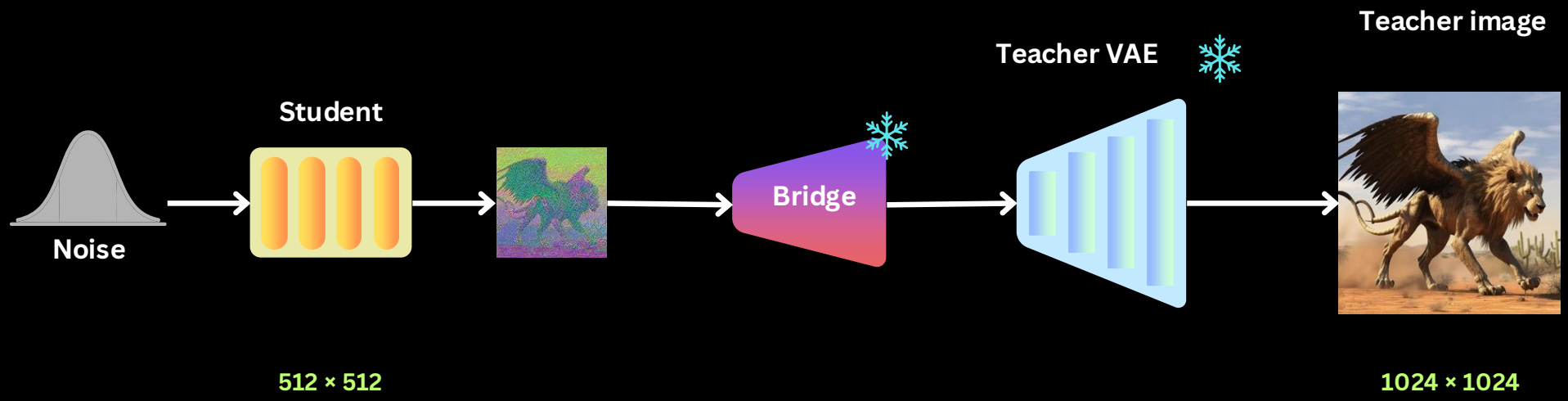
10.53
SD 1.5, from 9.49

9.75
SD 2.1, from 8.74

Three Teacher architecture families — U-Net, DiT, MM-DiT — in one 0.86 B Student.

Bonus: 1024 × 1024 from a Student trained at 512

The Bridge outputs a Teacher latent, so the Teacher's decoder can read it.



No additional training, still one function evaluation.

Conclusion

The teacher set is no longer fixed by the Student's latent space.

SAMPLE EXPERIMENT

5	≈ 5 M	1	5.37 → 10.53
Teachers, three architecture families	Bridge, outside the backbone	function evaluation	HPSv3 on SD 1.5

Teacher and Student can be chosen independently.

A deployed Student can be taught by a Teacher that arrives later.

arXiv:2606.32020
Qualcomm AI Research · UW–Madison · Johns Hopkins · Rutgers