Distilling Diffusion Models

Shreshth Saini · Sep 2026

shreshthsaini.github.io/blogs/diffusion-distillation.html



50 steps

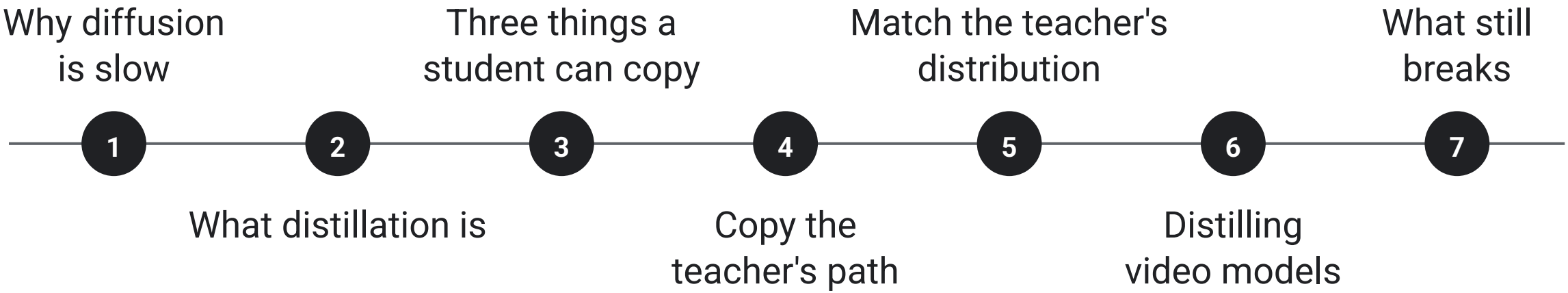


4 steps

[1] Yin et al. 2024, CC BY 4.0

What we will cover today

What is the student allowed to forget?



PART 1

Why diffusion is slow

Why does one image need a hundred network passes?

Sampling takes many small steps



the probability-flow ODE

$$\frac{dx}{dt} = v(x, t)$$

$$v(x, t) = \frac{x - \mathbb{E}[x_0 \mid x_t]}{t}$$

$$x_t = (1 - t)\, x_0 + t\, \varepsilon$$

cost per sample = NFE × network cost

[1] Yin et al. 2024, arXiv:2405.14867 · [49] Team Wan 2025, arXiv:2503.20314

From 100 network passes down to 4

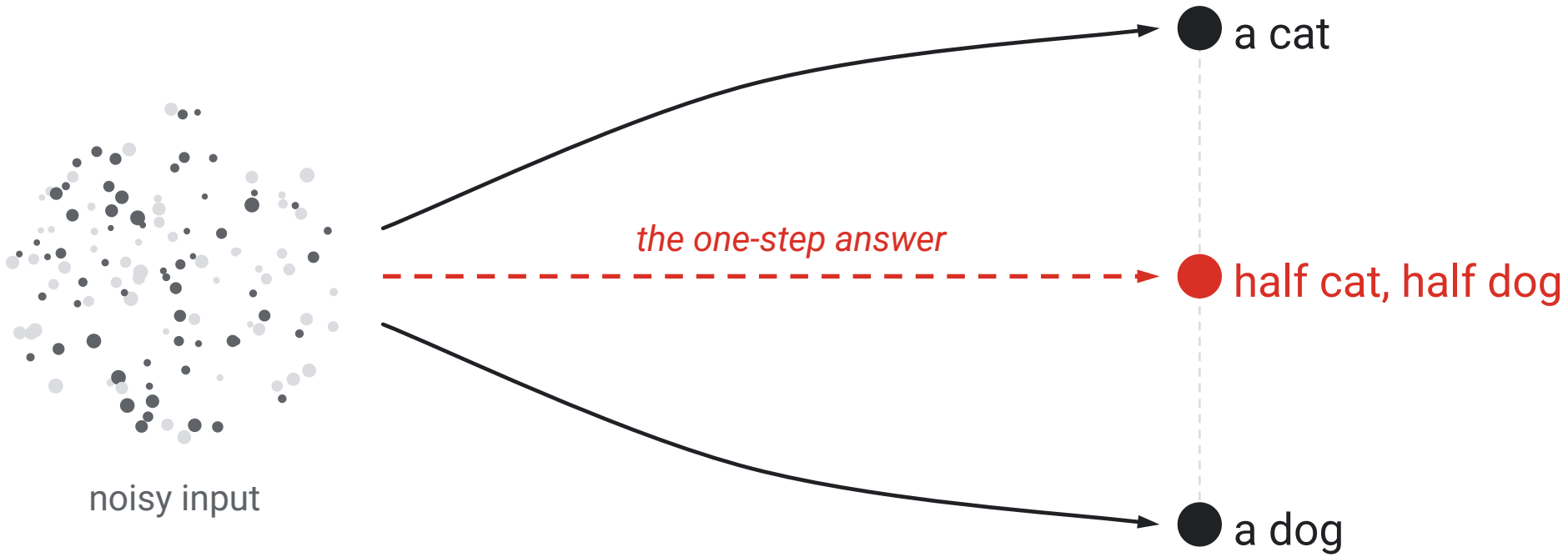


50 steps × 2 for guidance = 100 passes

Wan2.1 1.3B takes 103 s for one clip on an H100.

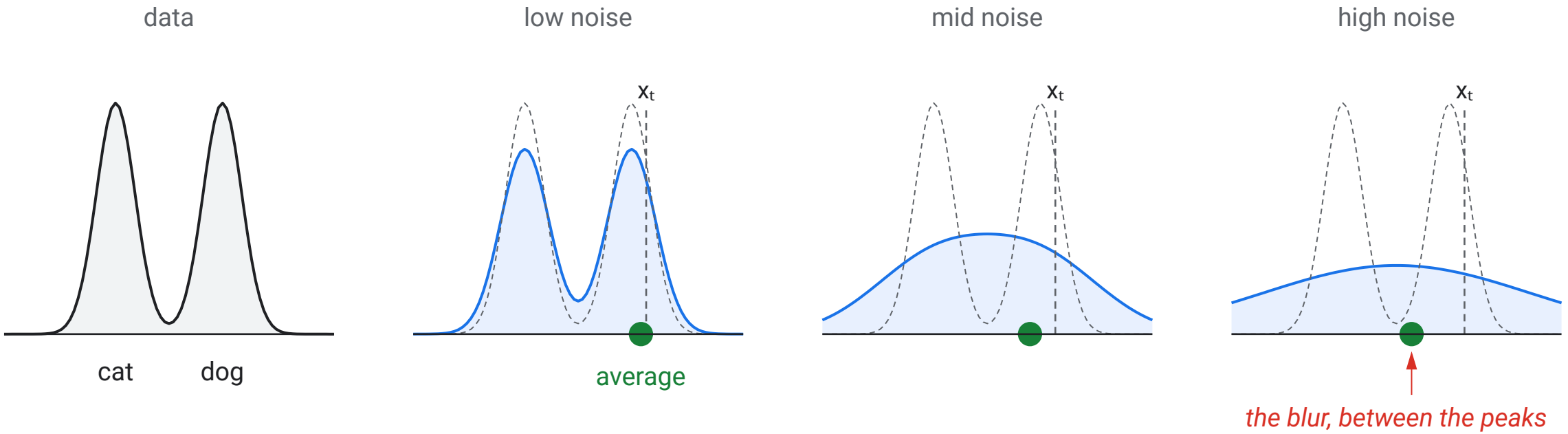
[1] Yin et al. 2024, arXiv:2405.14867 · [19] Huang et al. 2025, arXiv:2506.08009 · [49] Team Wan 2025, arXiv:2503.20314

Why one big step gives a blur



This noise could become a cat or a dog.
Asked to answer in one step, the network returns the average, a blur.

One step lands on the average image



$$\mathbb{E}[x_0 \mid x_t] = \frac{x_t + \sigma_t^2 \nabla_{x_t} \log p_t(x_t)}{a_t}$$

Squared-error training learns this average.

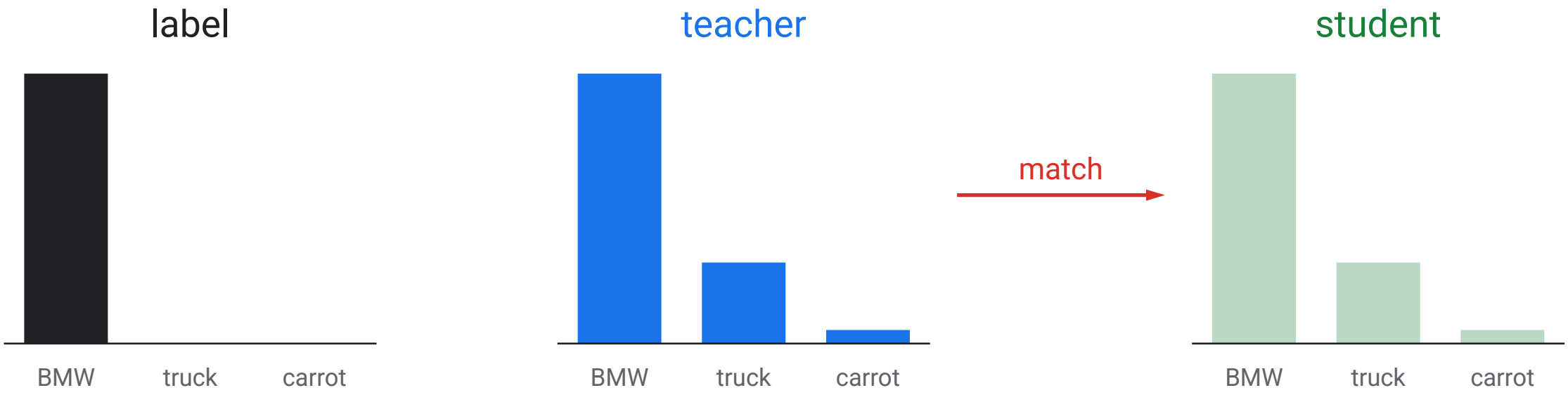
[2] Blau 2018 · [3] Robbins 1956 · [4] Efron 2011 · [5] Dieleman 2024

PART 2

What distillation is

How does a student learn from a teacher?

Distillation in one picture

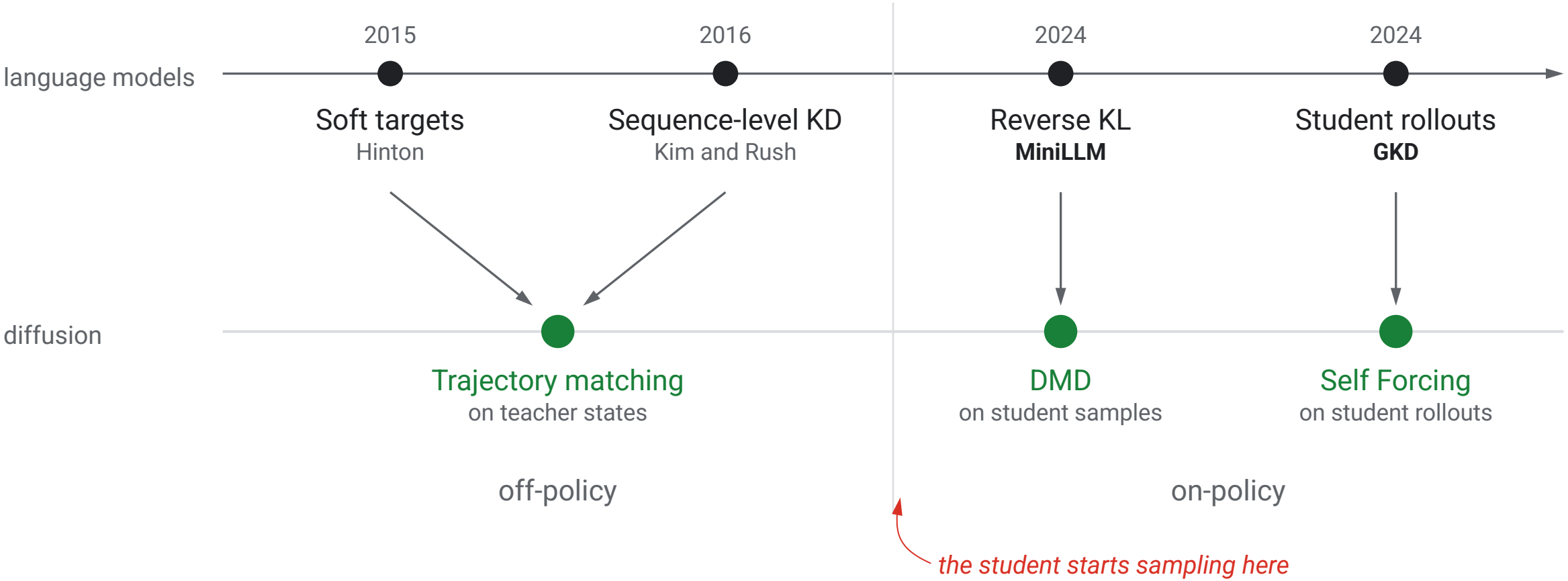


The label says BMW. The teacher says BMW, a bit garbage truck, almost no carrot.

The student learns from that extra shape.

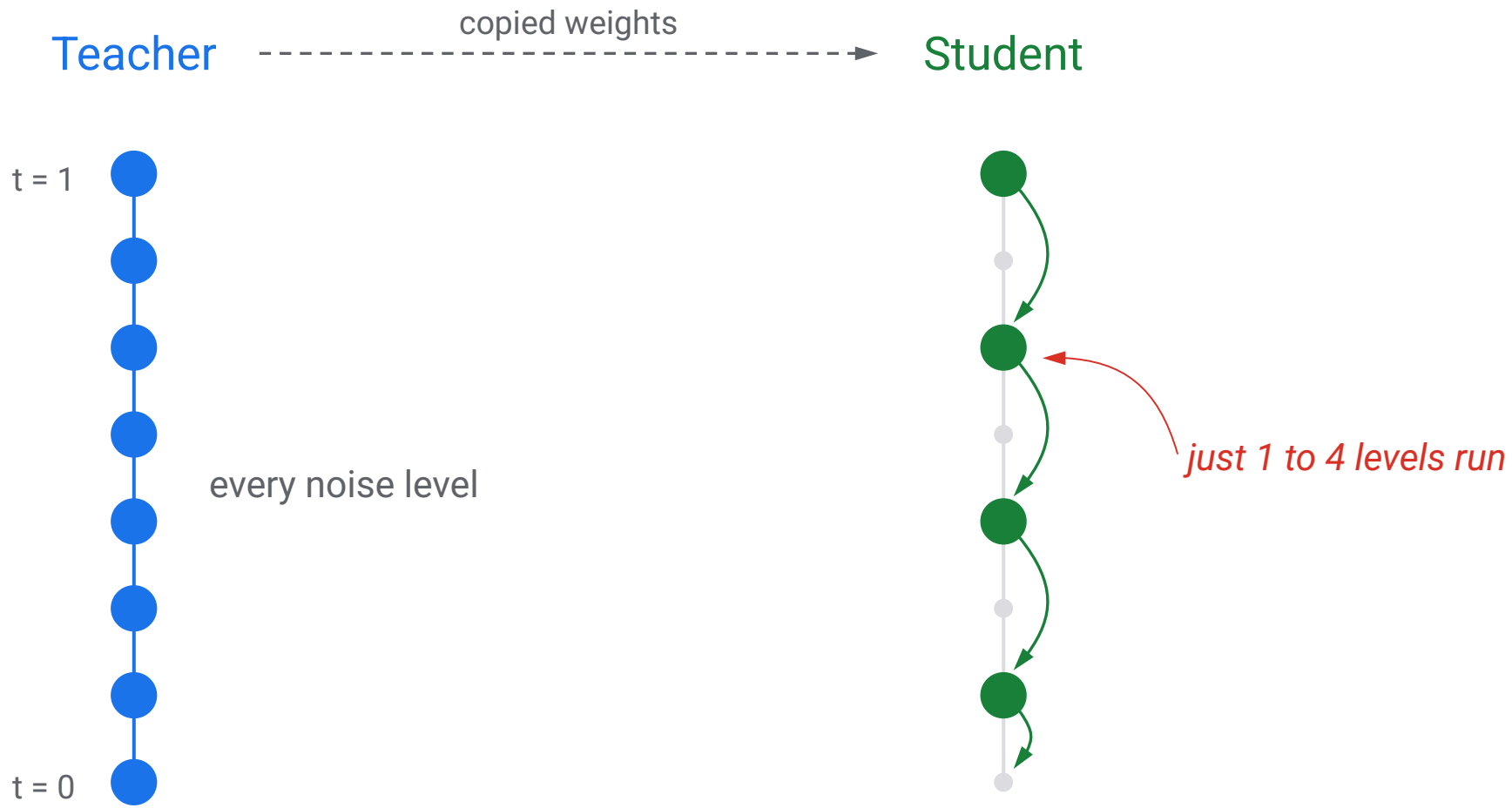
[7] Hinton et al. 2015, arXiv:1503.02531

Language models went through this first

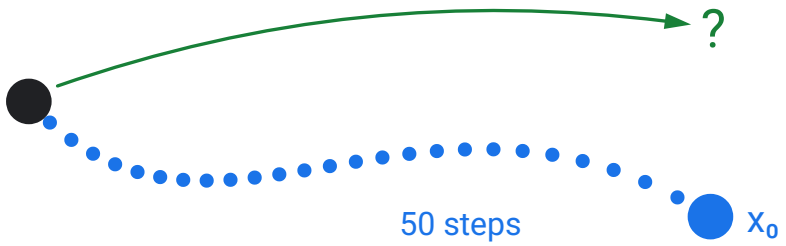


[7] Hinton 2015 · [8] Kim 2016 · [9] Gu 2024 · [10] Agarwal 2024 · [40] Yin 2024

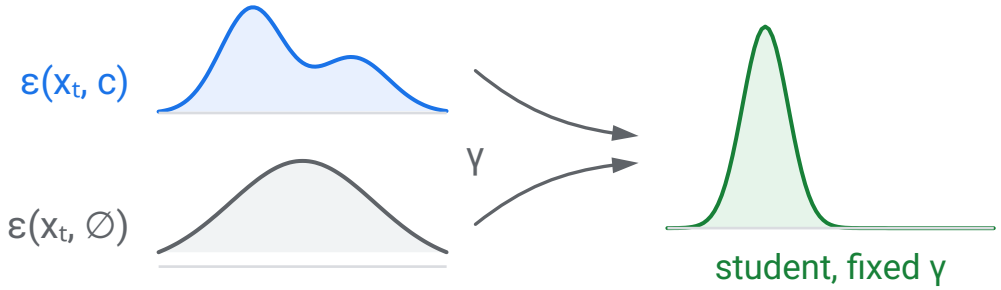
In diffusion, the student learns to skip steps



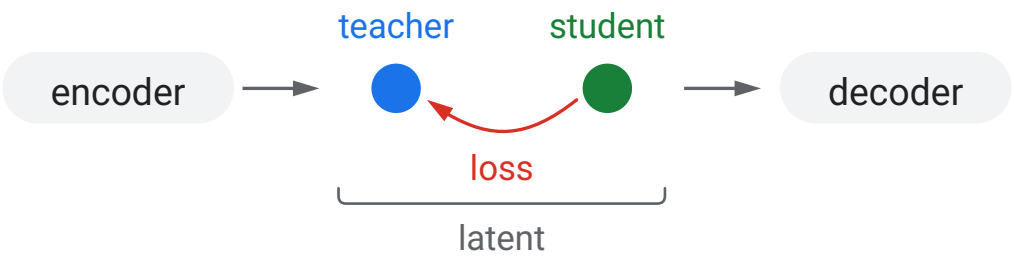
Why diffusion is harder to distill



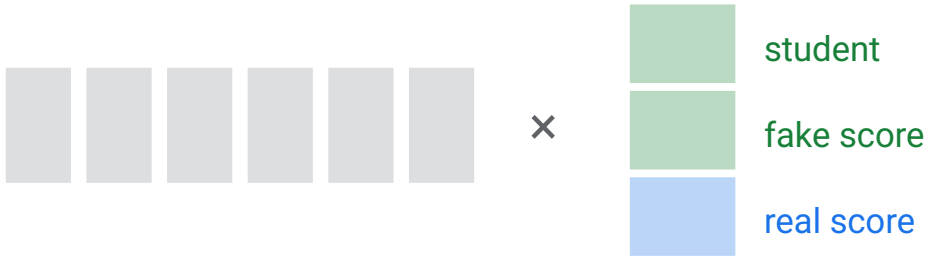
No few-step answer key



Guidance is in the target



The VAE sits outside



Video needs more memory

Who sets the target, and on what samples

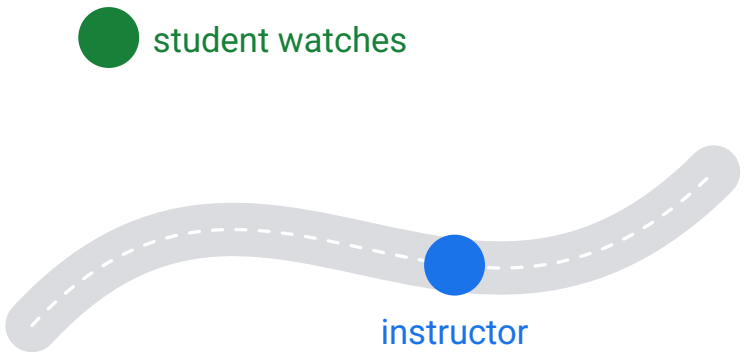
	Off-policy trained on data or teacher states	On-policy trained on the student's own samples
Teacher distillation target from a frozen teacher	• **progressive distillation** • **consistency distillation** • **LCM** • **reflow**	• **DMD2** • **SiD** • **LADD** • **Self Forcing**
Self-distillation target from its own output also called teacher-free	• consistency training • **shortcut models** • **MeanFlow**	*rare*

on-policy still has a teacher

[1] Yin 2024 · [6] Salimans 2022 · [19] Huang 2025 · [27] Song 2023 · [34] Geng 2025

Learning to drive, three ways

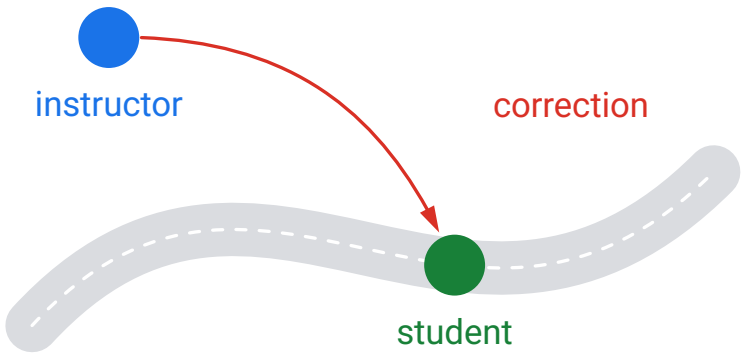
off-policy



Watch the instructor drive.

CD, **LCM**, **reflow**

on-policy

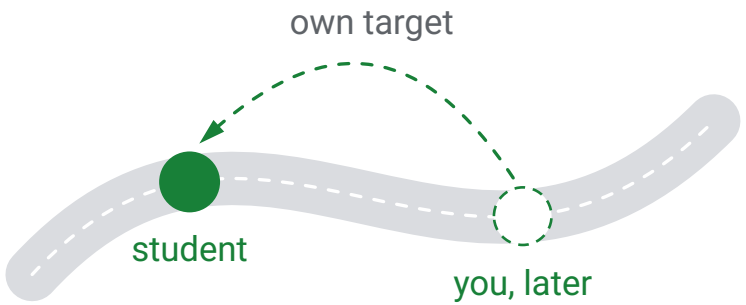


Drive while the instructor corrects you.

DMD, **ADD**, **Self Forcing**

still has a teacher

self-distillation

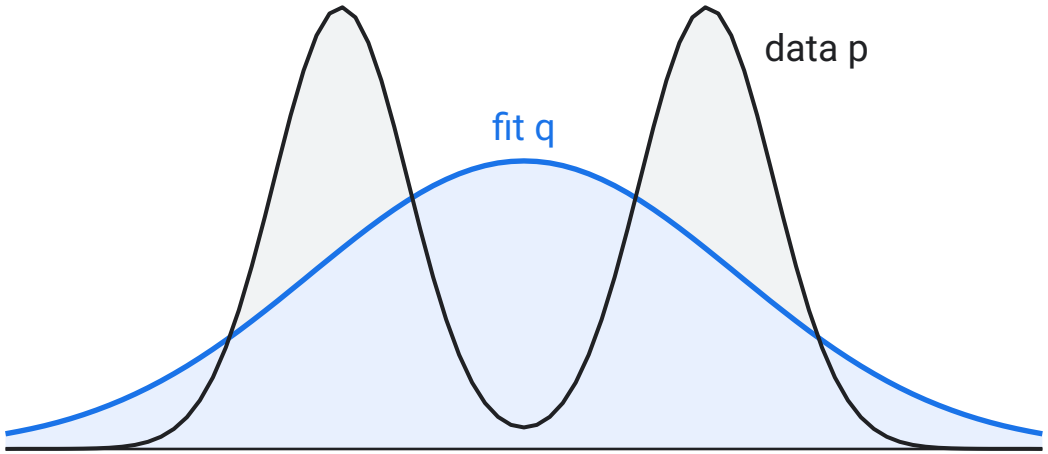


Or check your answer against your own later answer.

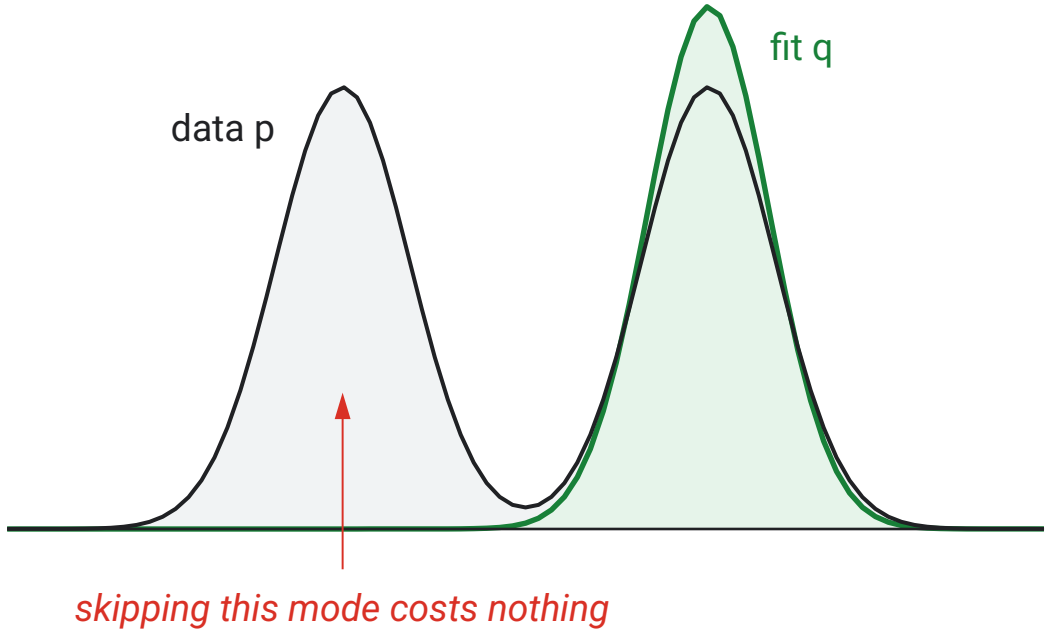
consistency training, **MeanFlow**

Forward KL spreads, reverse KL picks a peak

forward KL, the teacher's loss



reverse KL, the student's loss



A student can beat its teacher on FID while covering less.

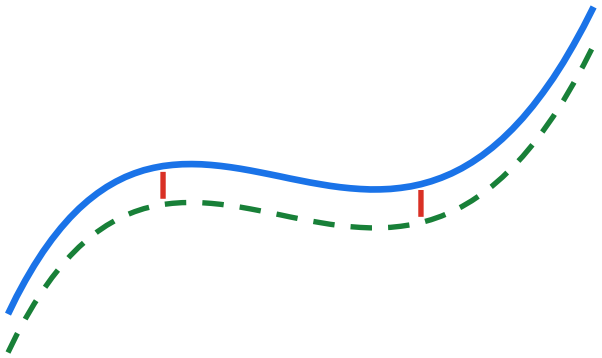
PART 3

Three things a student can copy

What exactly does the student try to match?

Three ways to learn to draw

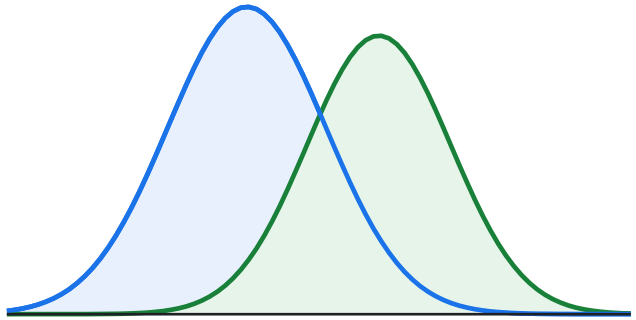
trace



Trace the teacher's strokes.

trajectory

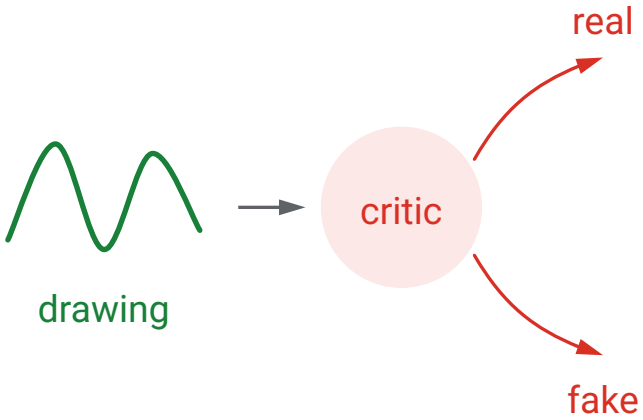
match the stack



Make your whole stack of drawings look like theirs.

distribution

critic

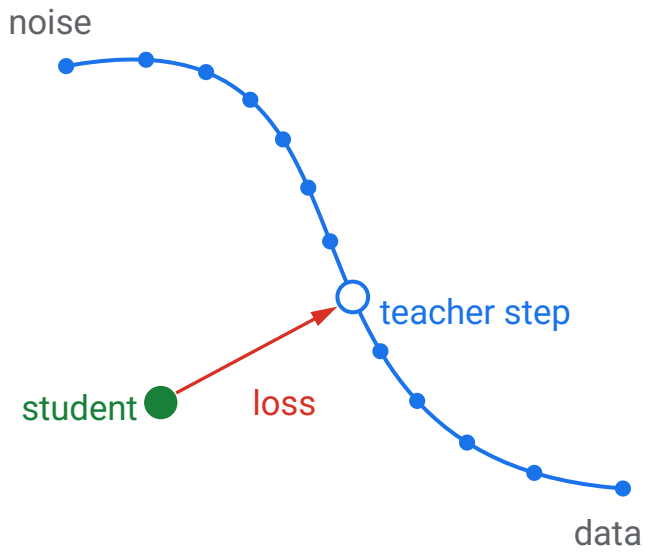


Or let a critic call each one real or fake.

discriminator

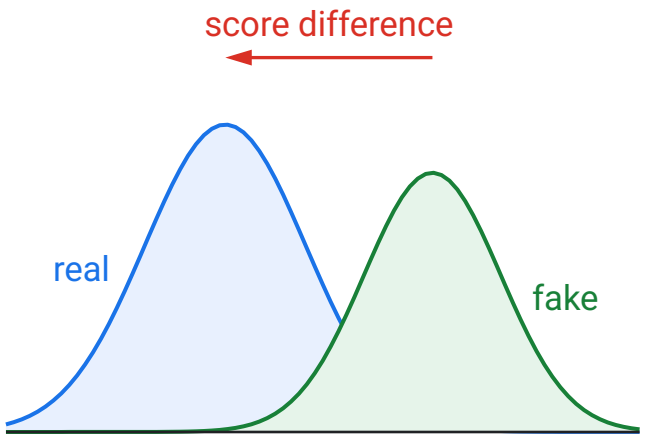
Every method copies one of three things

trajectory



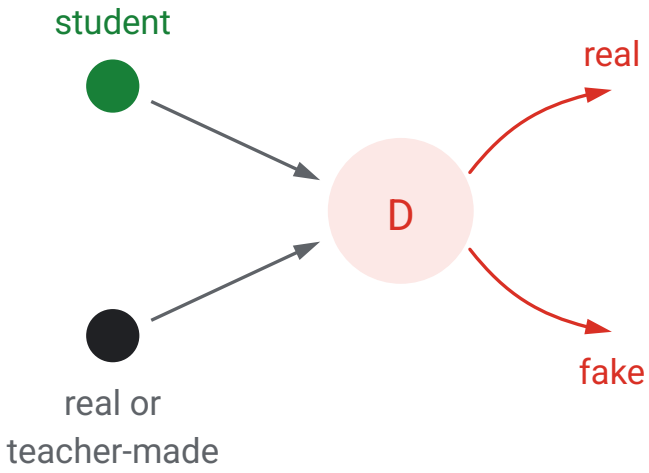
a point

distribution



a density

discriminator



a verdict

[27] Song et al. 2023, arXiv:2303.01469 · [40] Yin et al. 2024, arXiv:2311.18828 · [44] Sauer et al. 2023, arXiv:2311.17042

PART 4

Copy the teacher's path

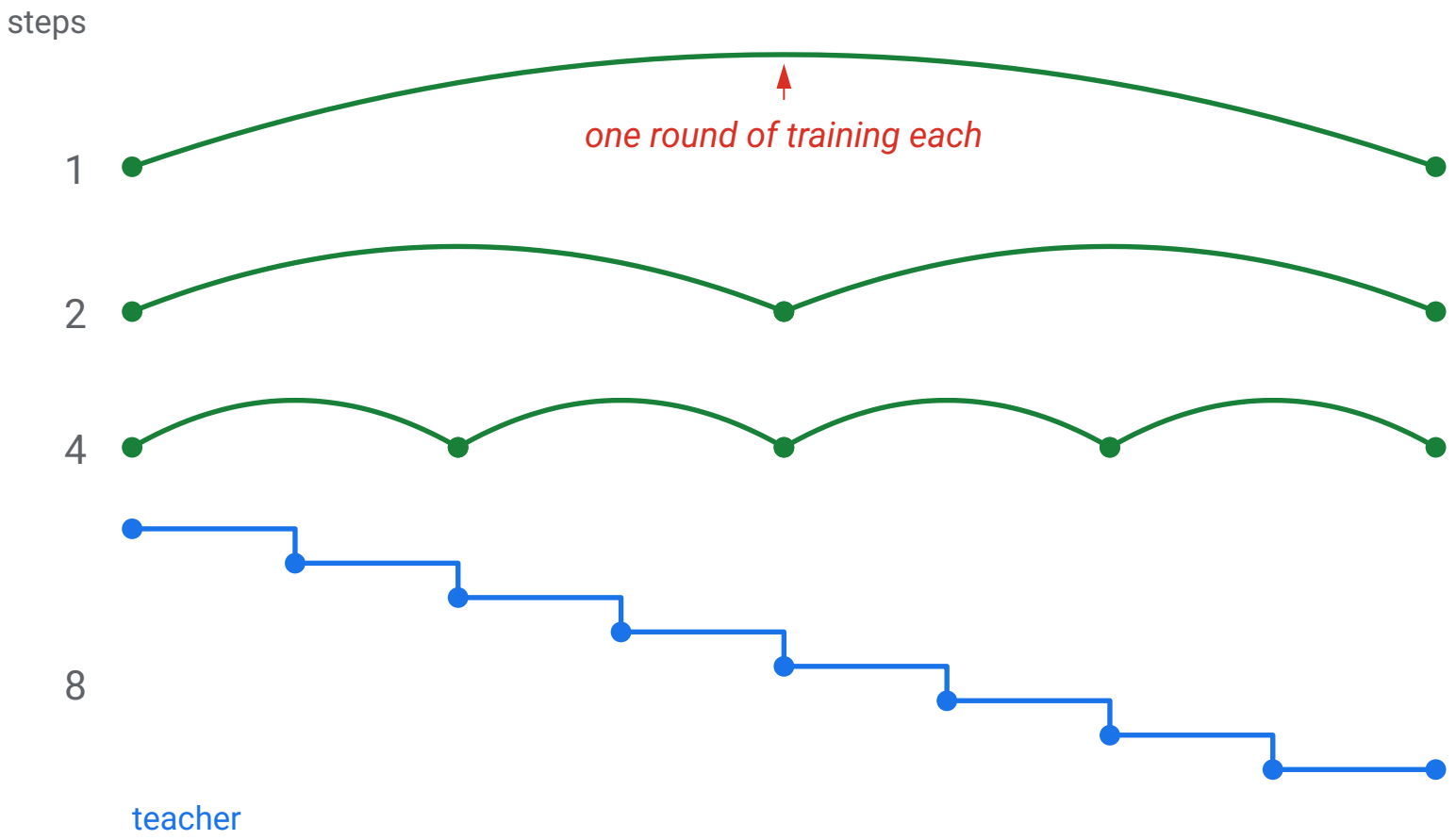
How did the first methods cut the steps?

Learning to skip stairs

Learn to cover two of the teacher's steps in one.

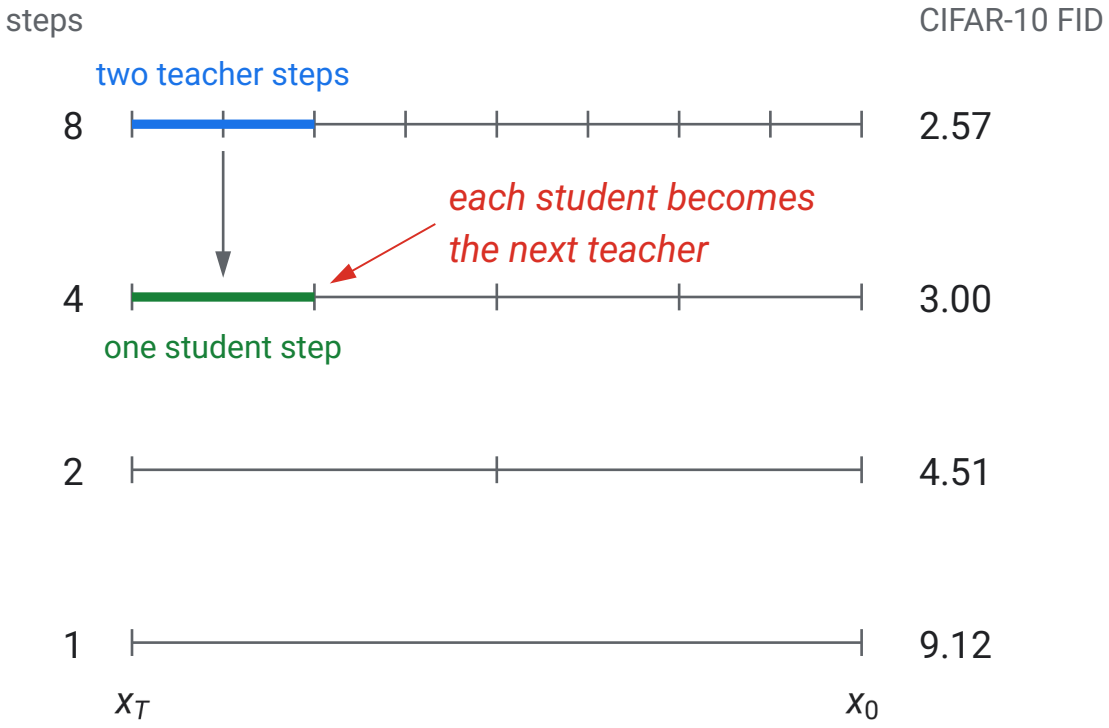
Then the student becomes the teacher.

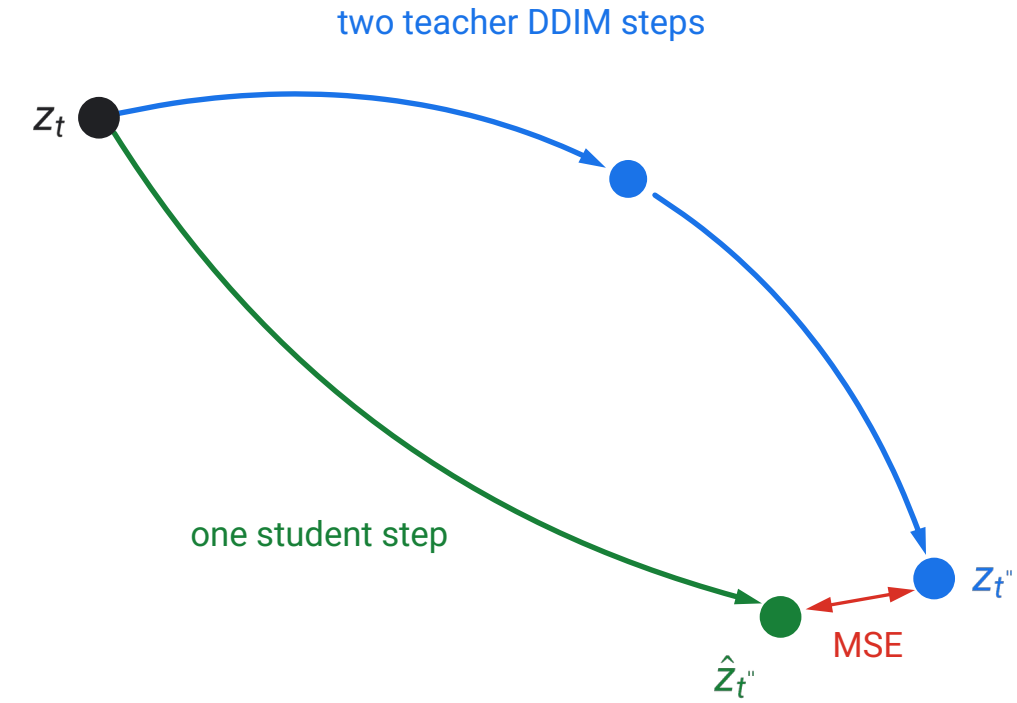
Repeat: 8, 4, 2, 1.



Progressive distillation halves the steps each round

teacher distillation, off-policy





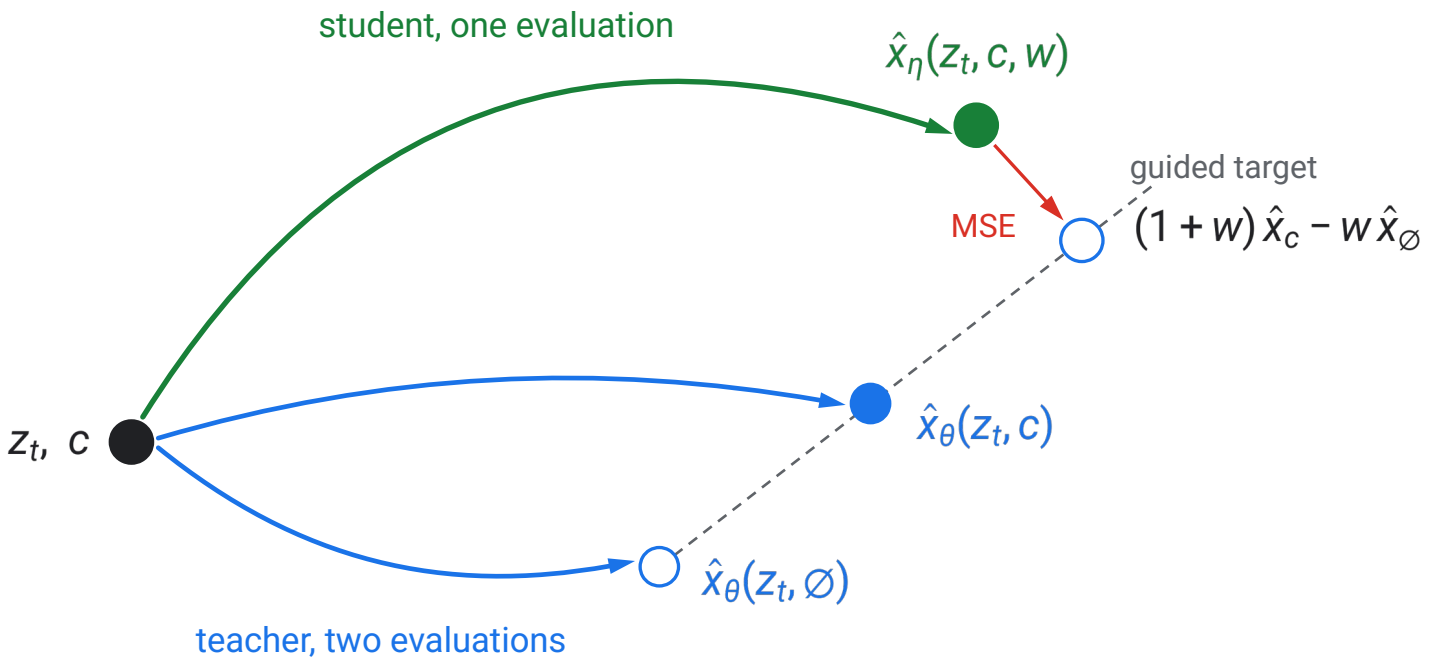
$$\tilde{x} = \frac{z_{t''} - (\sigma_{t''}/\sigma_t)\, z_t}{\alpha_{t''} - (\sigma_{t''}/\sigma_t)\, \alpha_t}$$

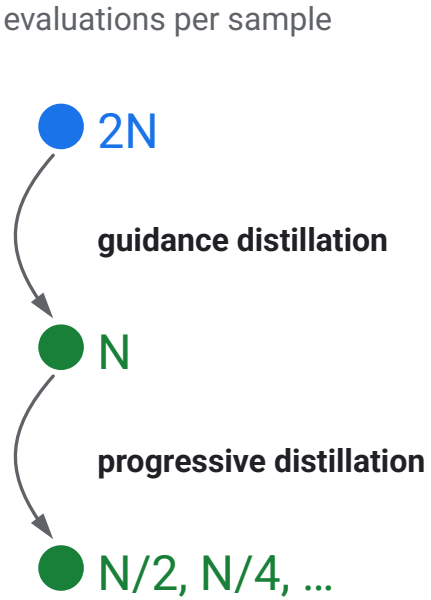
$$\mathscr{L}_\theta = w(\lambda_t)\, \|\tilde{x} - \hat{x}_\theta(z_t)\|_2^2$$

[6] Salimans and Ho 2022, arXiv:2202.00512

Guidance distillation folds two passes into one

teacher distillation, off-policy





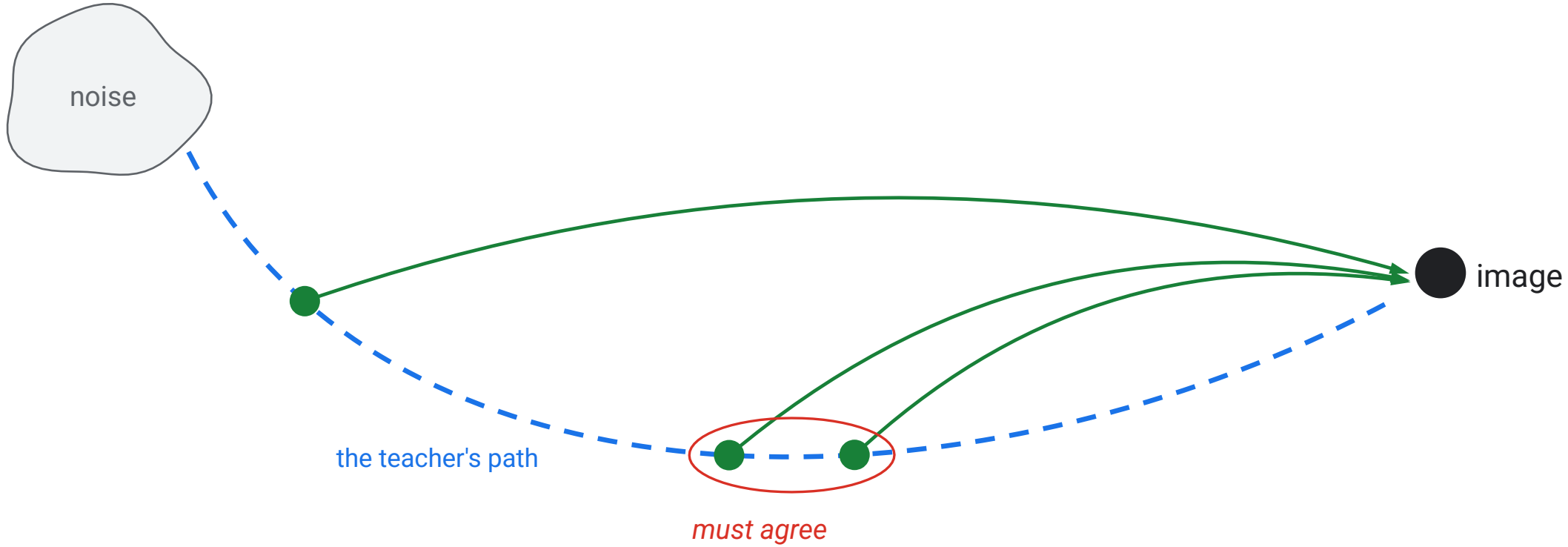
FID: guided teacher 9.84, one-step student 7.54 ← *the student beats its teacher*

ImageNet 64, w = 1

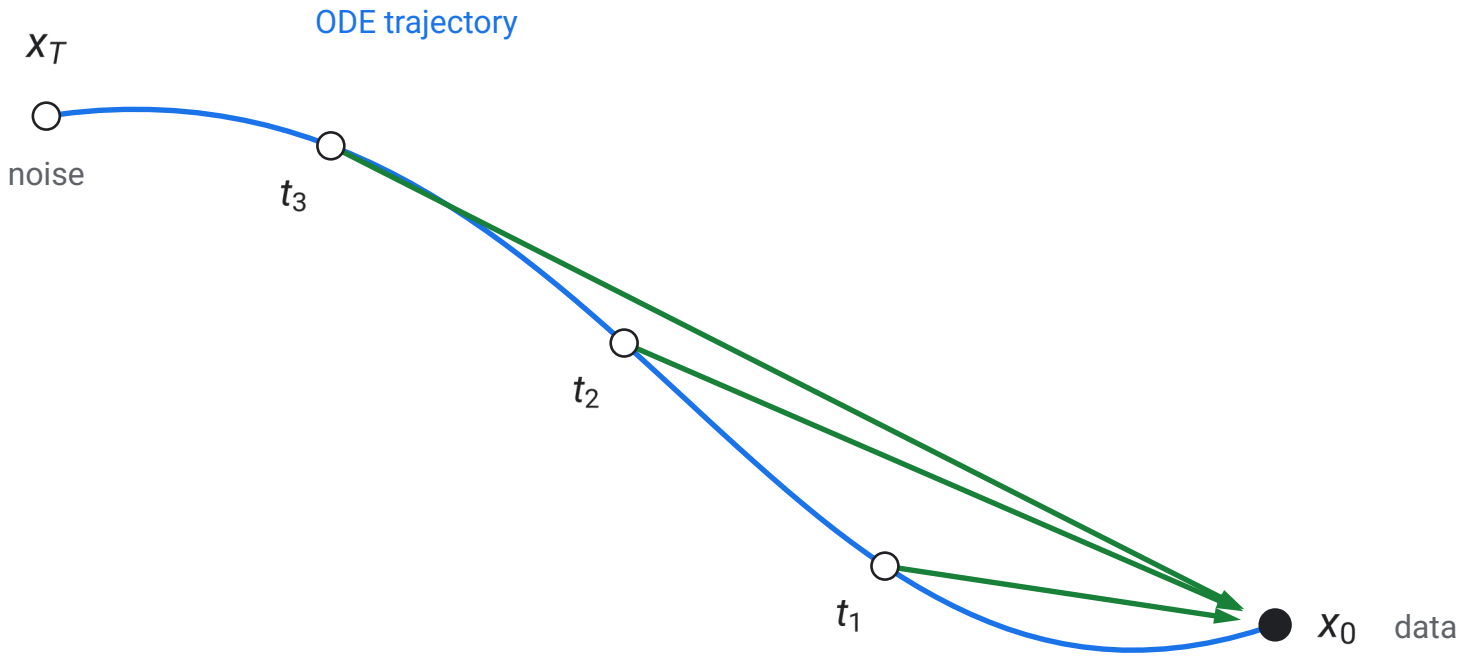
[20] Meng et al. 2023, arXiv:2210.03142

The idea behind consistency models

From any point on the trip, say where it ends.
Two nearby points must give the same answer.



A consistency model jumps straight to the image



$$f_\theta(x_t, t) = f_\theta(x_s, s)$$

same trajectory, same output

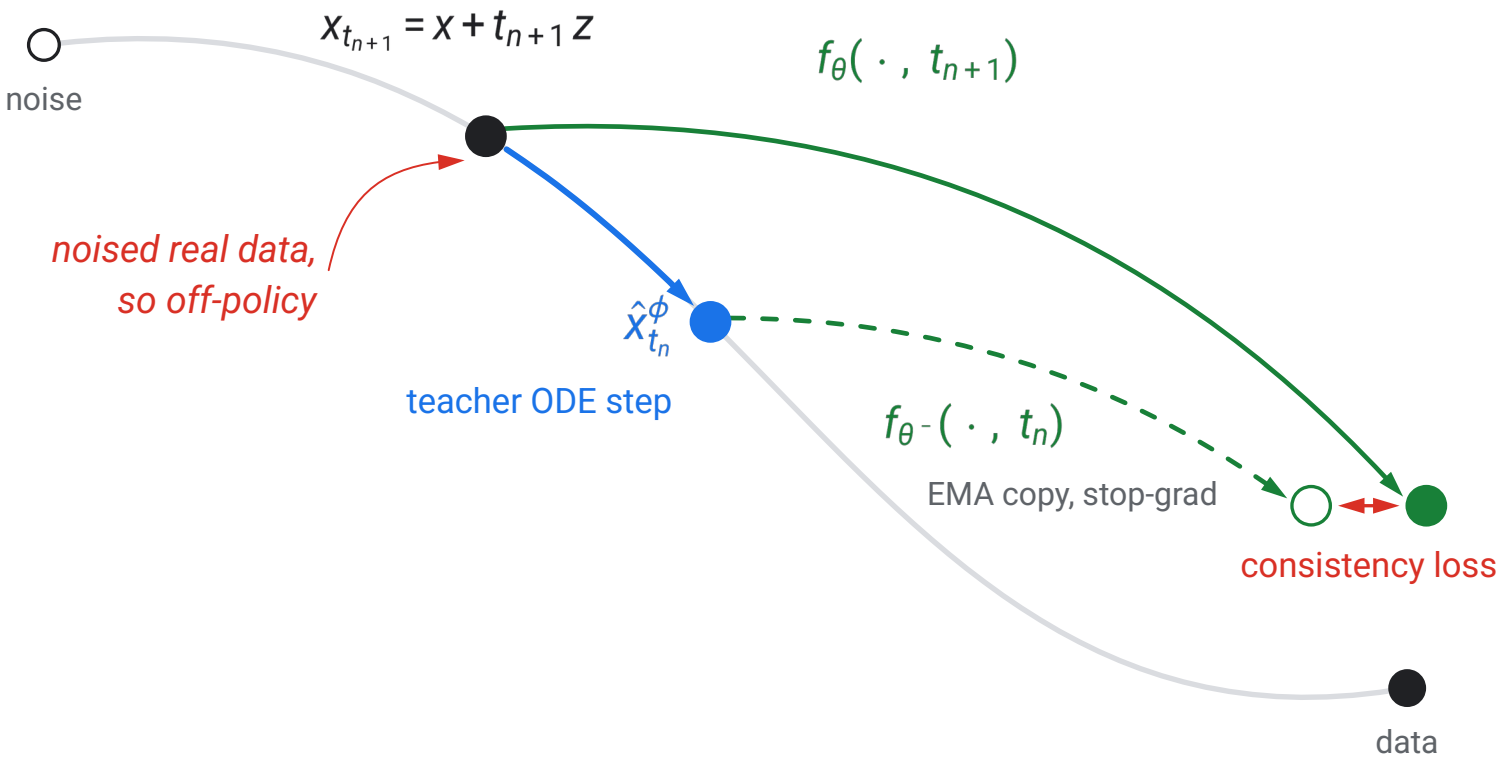
$$f_\theta(x, t) = c_{\text{skip}}(t)\, x + c_{\text{out}}(t)\, F_\theta(x, t)$$

$$c_{\text{skip}}(\varepsilon) = 1, \quad c_{\text{out}}(\varepsilon) = 0$$

$$f(x_\varepsilon, \varepsilon) = x_\varepsilon \longleftarrow \textit{identity here, by construction}$$

Consistency distillation makes neighbours agree

teacher distillation, off-policy



FID	1 step	2 steps	teacher
CD, CIFAR-10	3.55	2.93	2.04
CD, ImageNet 64	6.20	4.70	2.44

$$\mathscr{L}_{\text{CD}} = \mathbb{E}\left[\lambda(t_n)\, d\left(f_\theta(x_{t_{n+1}}, t_{n+1}),\ f_{\theta^-}(\hat{x}^{\phi}_{t_n}, t_n)\right)\right]$$

Consistency training drops the teacher

self-distillation (teacher-free), off-policy

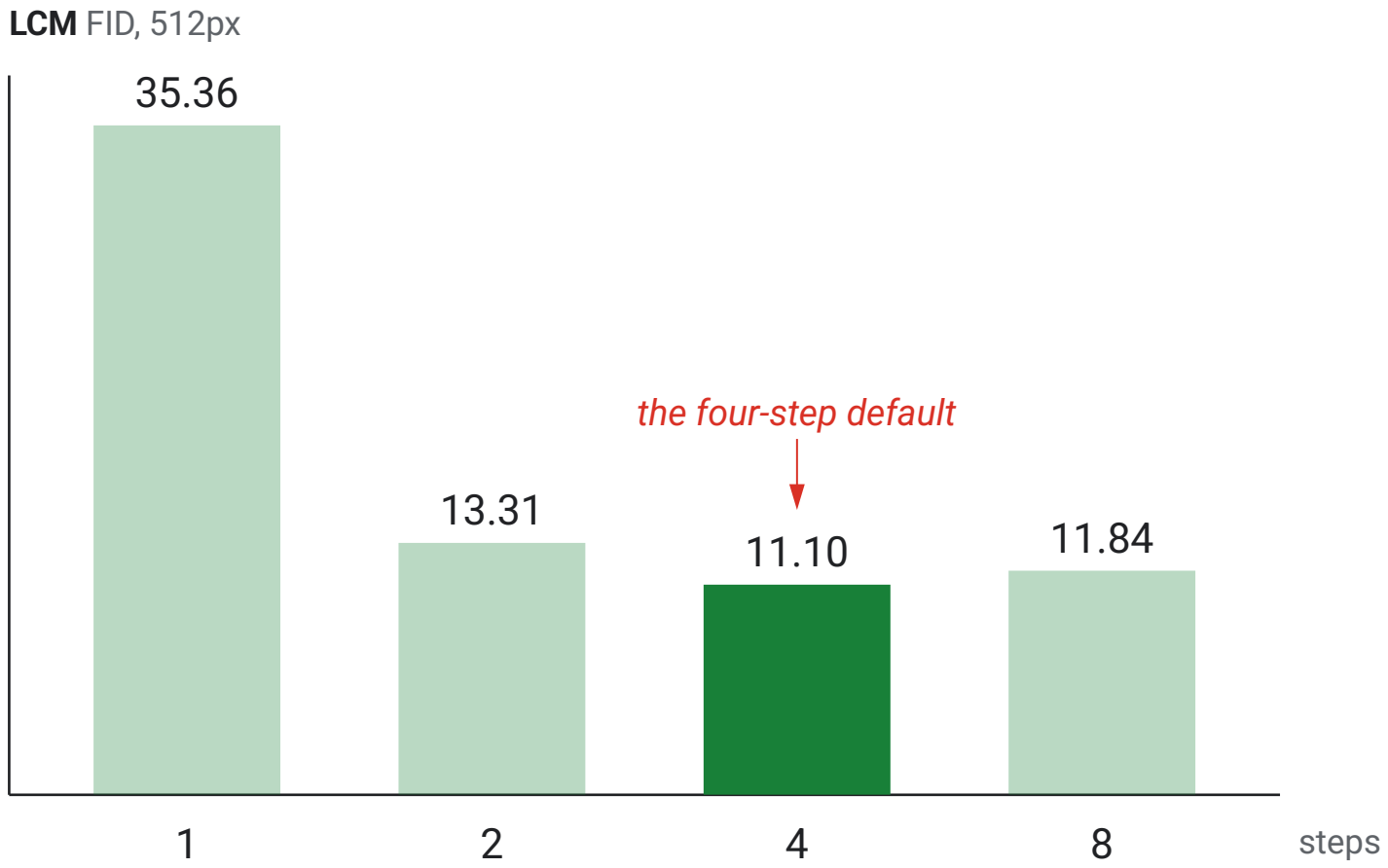
Method	Teacher	Benchmark	1-step FID	2-step FID
CD, 2023	EDM	CIFAR-10	3.55	2.93
CT, 2023	none	CIFAR-10	8.70	5.83
iCT-deep, 2024	none	CIFAR-10	2.51	2.24
sCT-XXL, 2025	none	ImageNet 512	4.29	3.76

no teacher, and it beats CD

[27] Song et al. 2023, arXiv:2303.01469 · [28] Song and Dhariwal 2024, arXiv:2310.14189 · [30] Lu and Song 2025, arXiv:2410.11081

LCM brought four steps to Stable Diffusion

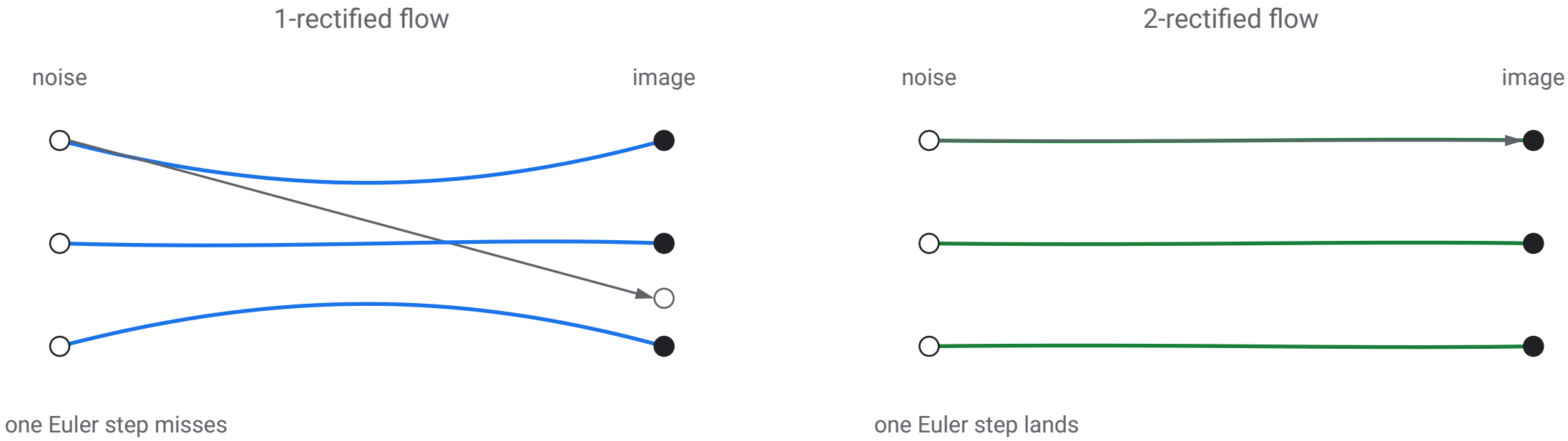
teacher distillation, off-policy



[24] Luo et al. 2023, arXiv:2310.04378

Reflow straightens the path

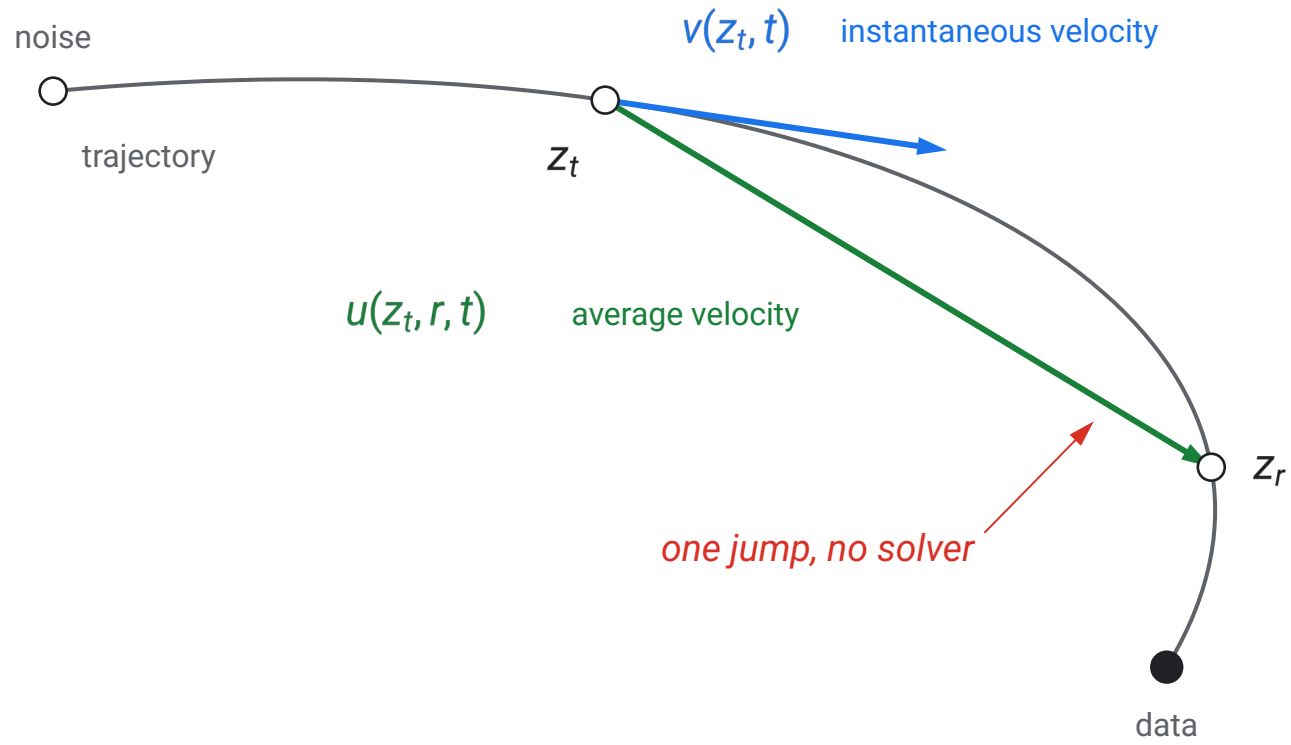
teacher distillation, off-policy



On SD 1.5, **InstaFlow** gets FID 23.4 in one step, against the teacher's 20.1. ← *the price of a straight path*

Flow maps learn whole jumps at once

self-distillation, off-policy



$$z_r = z_t - (t - r)\, u(z_t, r, t)$$

Shortcut models make one 2d step match two of their own d steps.

$$u(z_t, r, t) = v(z_t, t) - (t - r)\frac{d}{dt} u(z_t, r, t)$$

MeanFlow identity

[34] Geng et al. 2025, arXiv:2505.13447 · [33] Frans et al. 2025, arXiv:2410.12557

The trouble with copying the path



one step blurs ← *Part 5 starts here*



every teacher error is copied

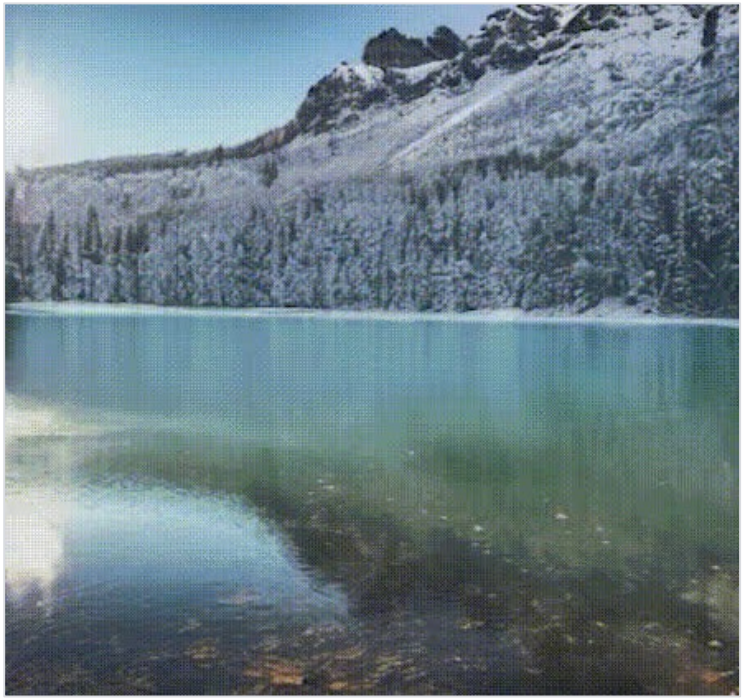
the teacher is the ceiling



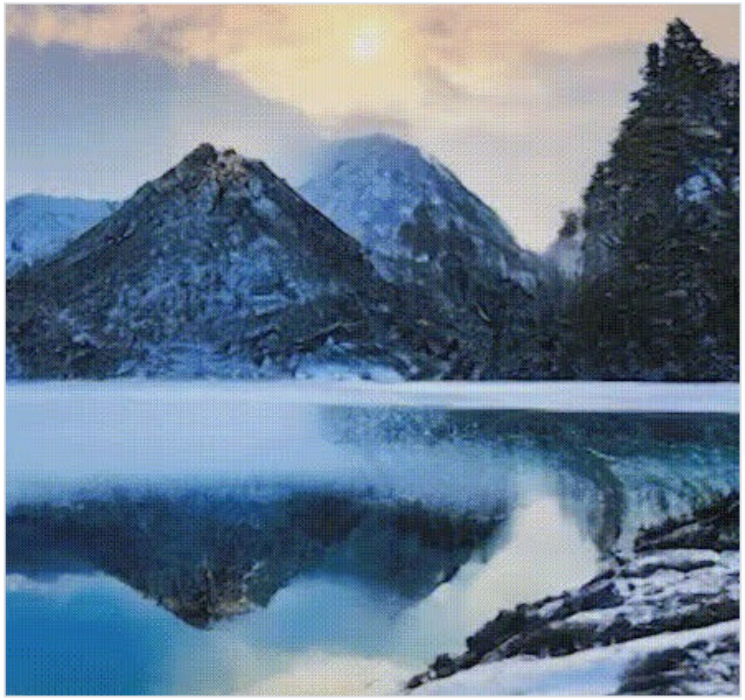
keeps the noise-to-image pairing and variety

One step on a curved path blurs

straightened by **reflow**



SD 1.5, 25 steps



InstaFlow, 1 step

Progressive distillation, curved path



256 steps

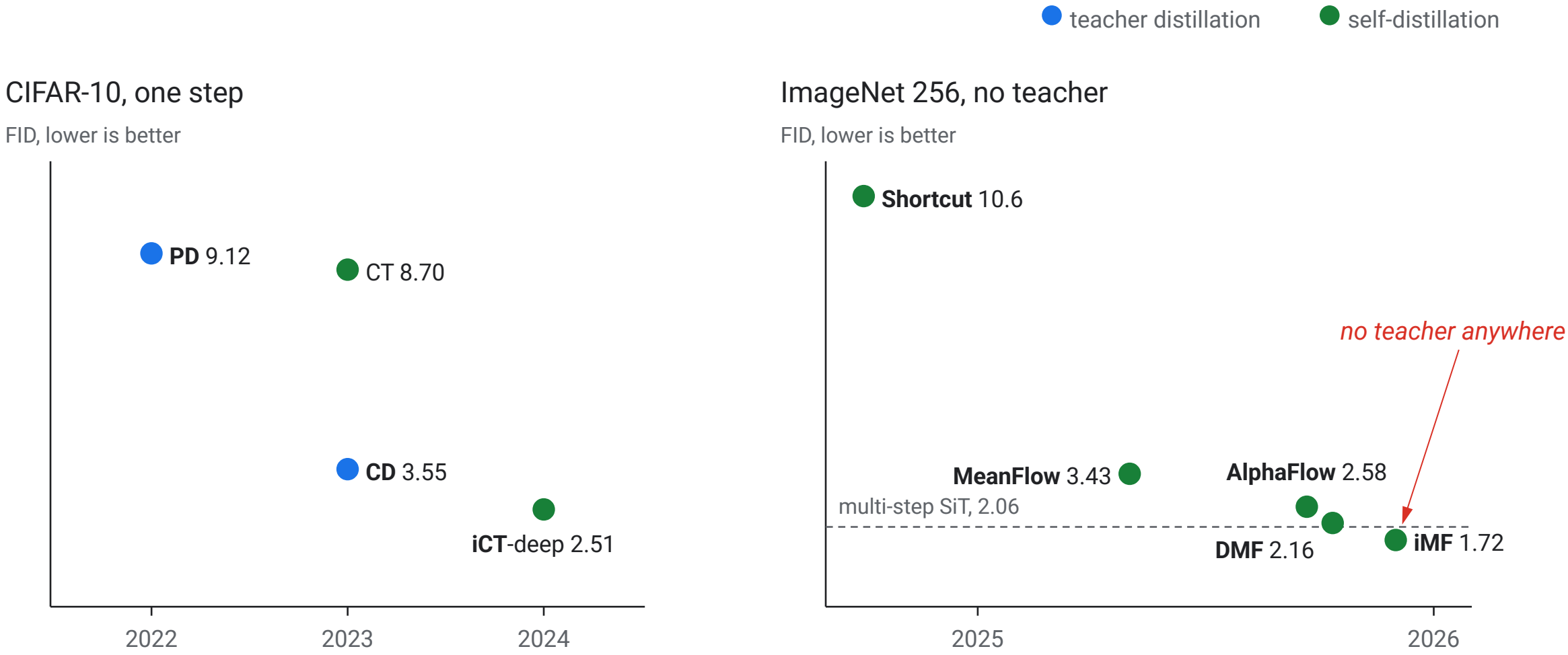


1 step

the average image, drawn

[32] Liu et al. 2024, arXiv:2309.06380 (MIT) · [6] Salimans and Ho 2022, arXiv:2202.00512, Fig. 3 · cropped

One-step scores kept getting better



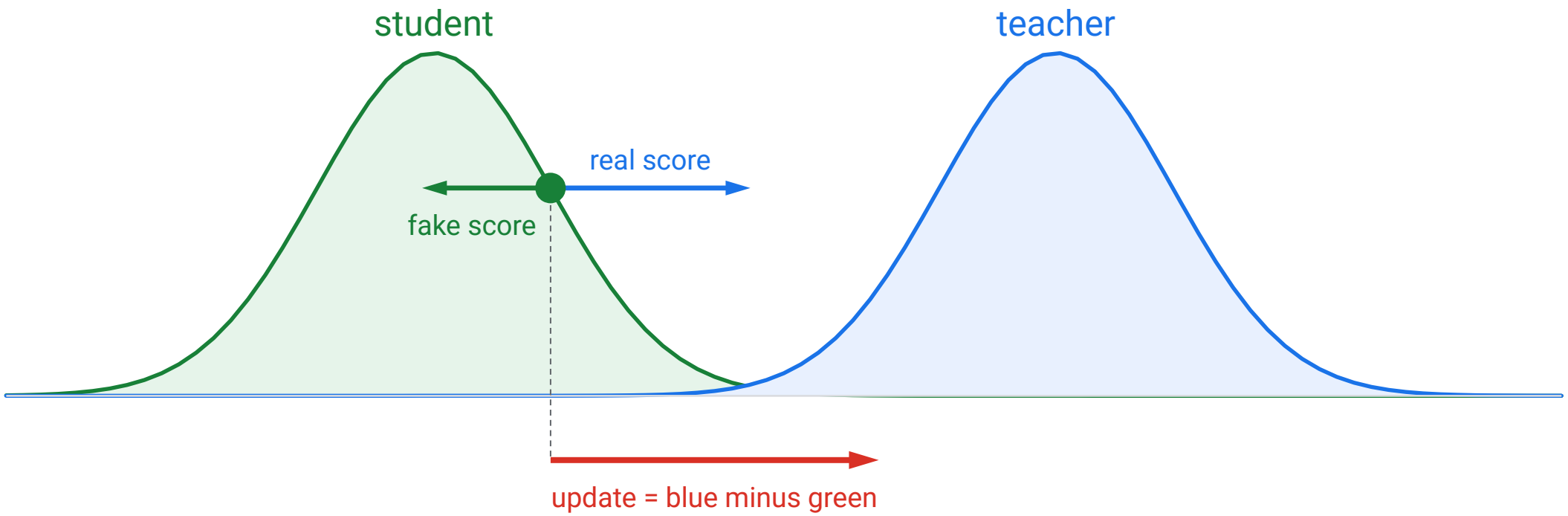
PART 5

Match the teacher's distribution

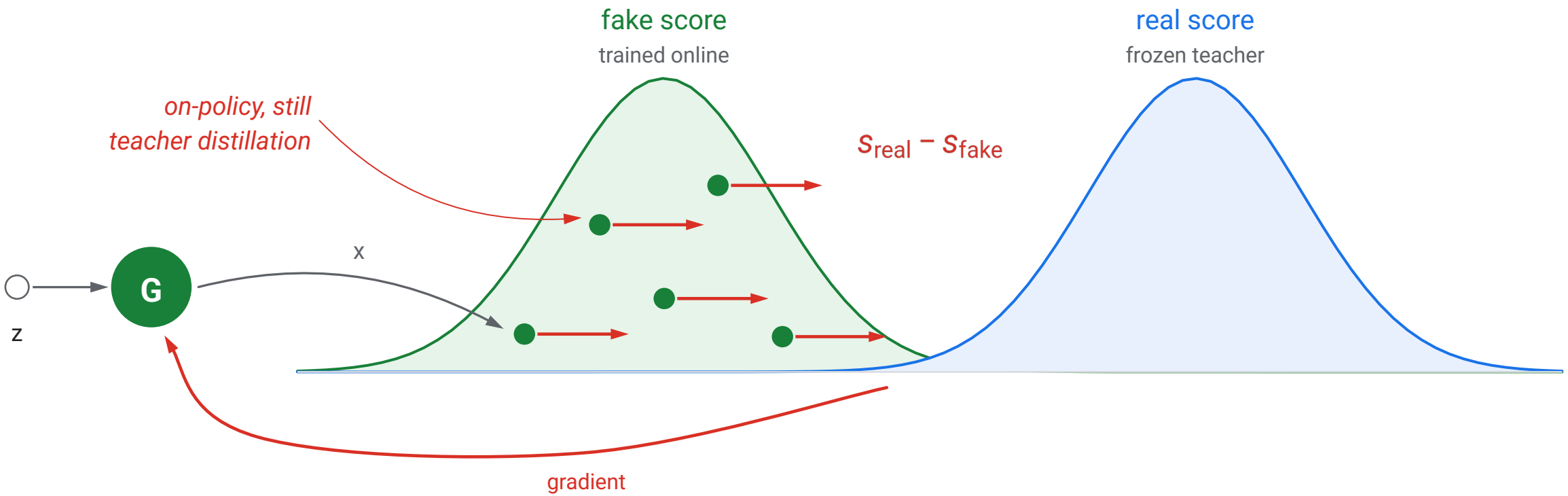
What if we judge the result and ignore the path?

Distribution matching in one picture

Push each student sample toward where the teacher's samples live, and away from where the student's own samples pile up.



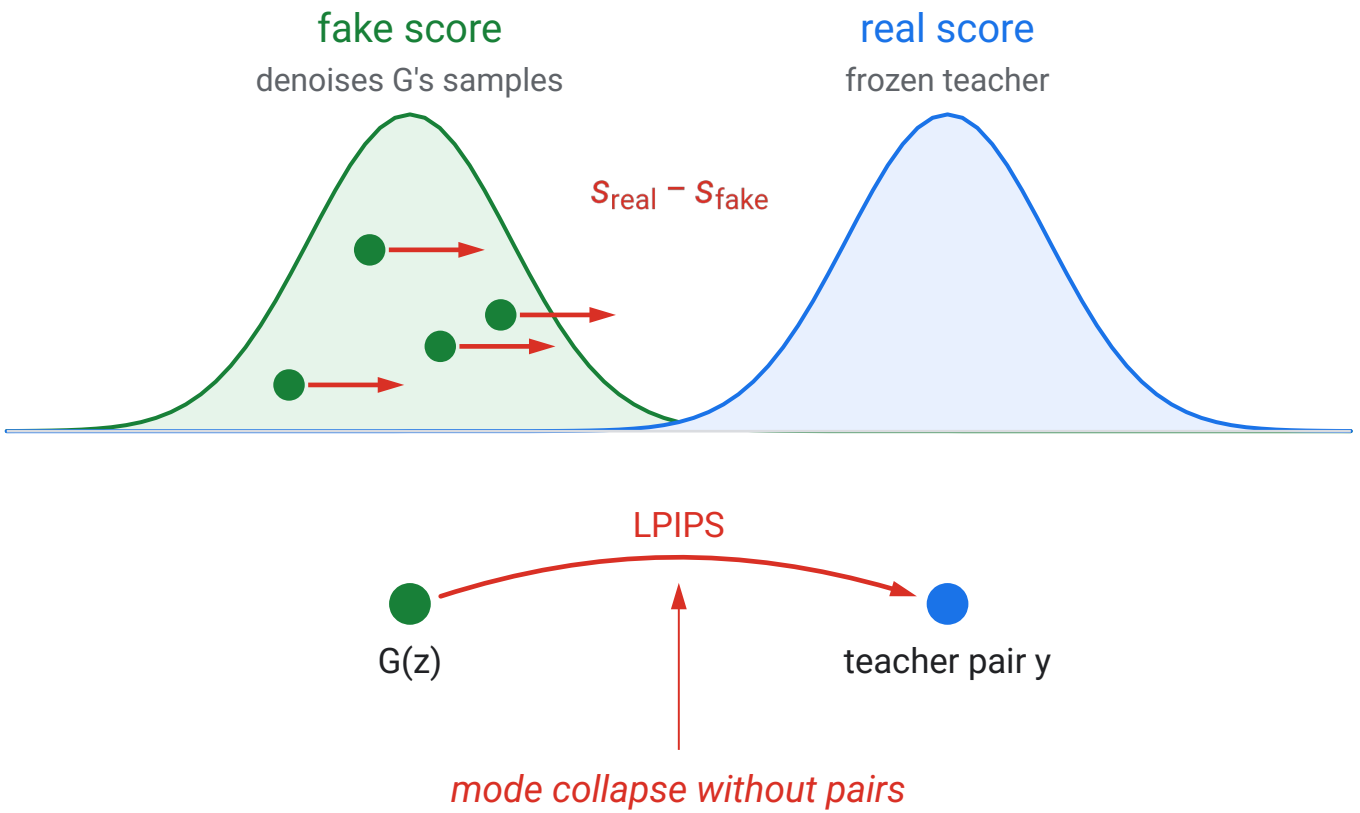
DMD moves samples by a score difference



$$\nabla_\theta D_{KL}(p_{fake} \,||\, p_{real}) \simeq \mathbb{E}_{z,t,x,x_t}\left[w_t\, \alpha_t\, (s_{fake}(x_t, t) - s_{real}(x_t, t))\, \frac{dG}{d\theta}\right]$$

[40] Yin et al. 2024, arXiv:2311.18828

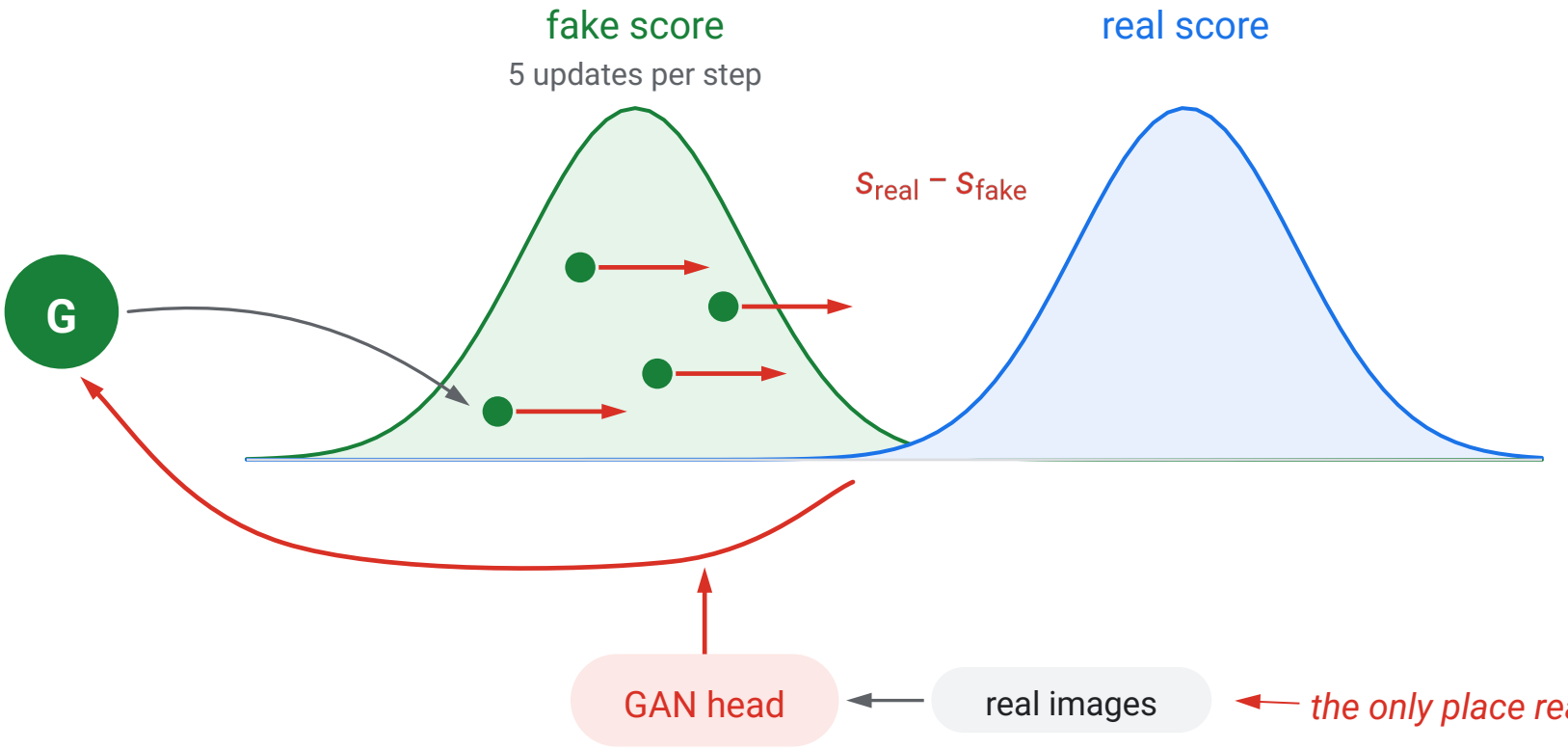
DMD also needed teacher pairs



One step, FID	CIFAR-10	ImageNet 64
regression only	3.82	9.21
distribution only	5.58	5.61
both	**2.66**	**2.62**

[40] Yin et al. 2024, arXiv:2311.18828

DMD2 drops the pairs and adds a GAN



ImageNet 64, one step	FID
DMD	2.62
drop pairs	3.48
+ 5 critic updates	2.61
+ GAN head	**1.51**
GAN only	2.52

[1] Yin et al. 2024, arXiv:2405.14867

Four DMD2 steps next to fifty SDXL steps

SDXL, 50 steps, guidance

DMD2, 4 steps







50 steps



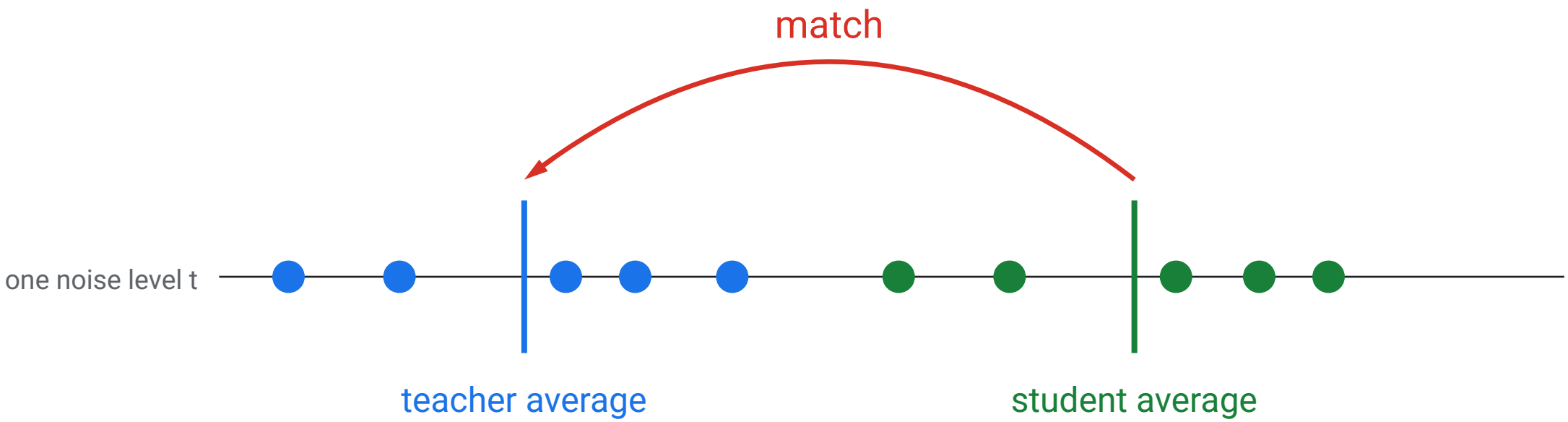
4 steps

FID 19.32 vs 19.36 · LPIPS diversity 0.61 vs 0.64

[1] Yin et al. 2024, arXiv:2405.14867, Figs. 6 and 10 · images CC BY 4.0, cropped

Moment matching in one picture

Do not check each sample. At every noise level, make the student's average guess of the clean image match the teacher's.



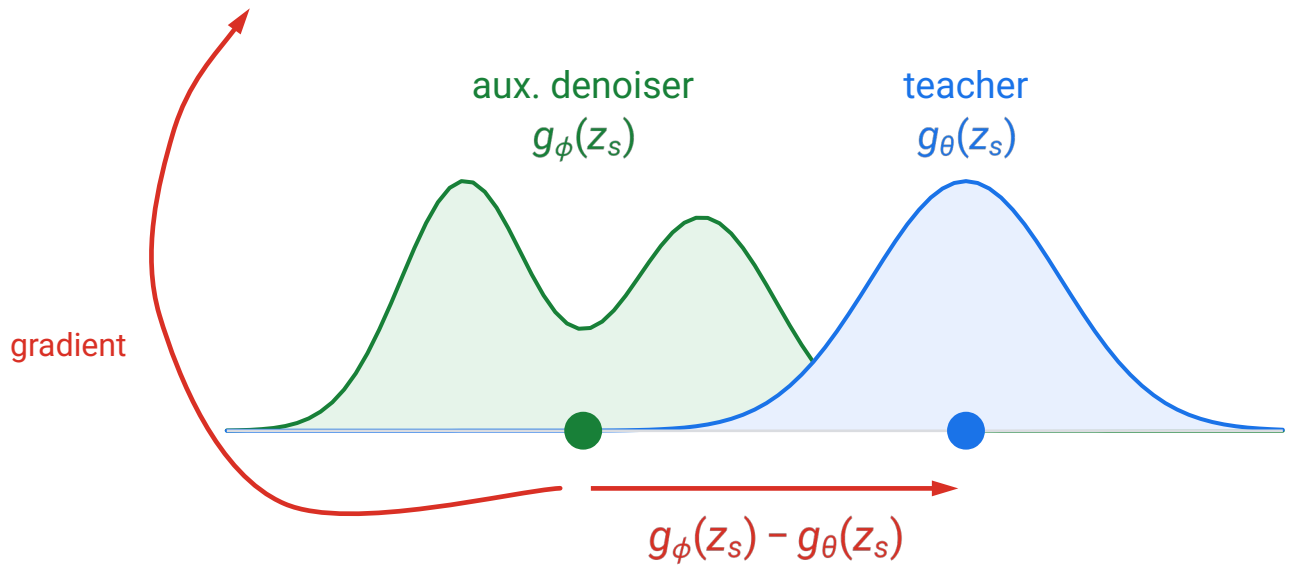
Moment matching compares average predictions

1 real image
+ noise → z_t

2 student
$\tilde{x} = g_\eta(z_t)$

3 sampler step
$z_s \sim q(z_s \mid z_t, \tilde{x})$

starts from noised real data



$$\mathbb{E}_g[\tilde{x} \mid z_s] = \mathbb{E}_q[x \mid z_s]$$

matched for every z_s

ImageNet 64

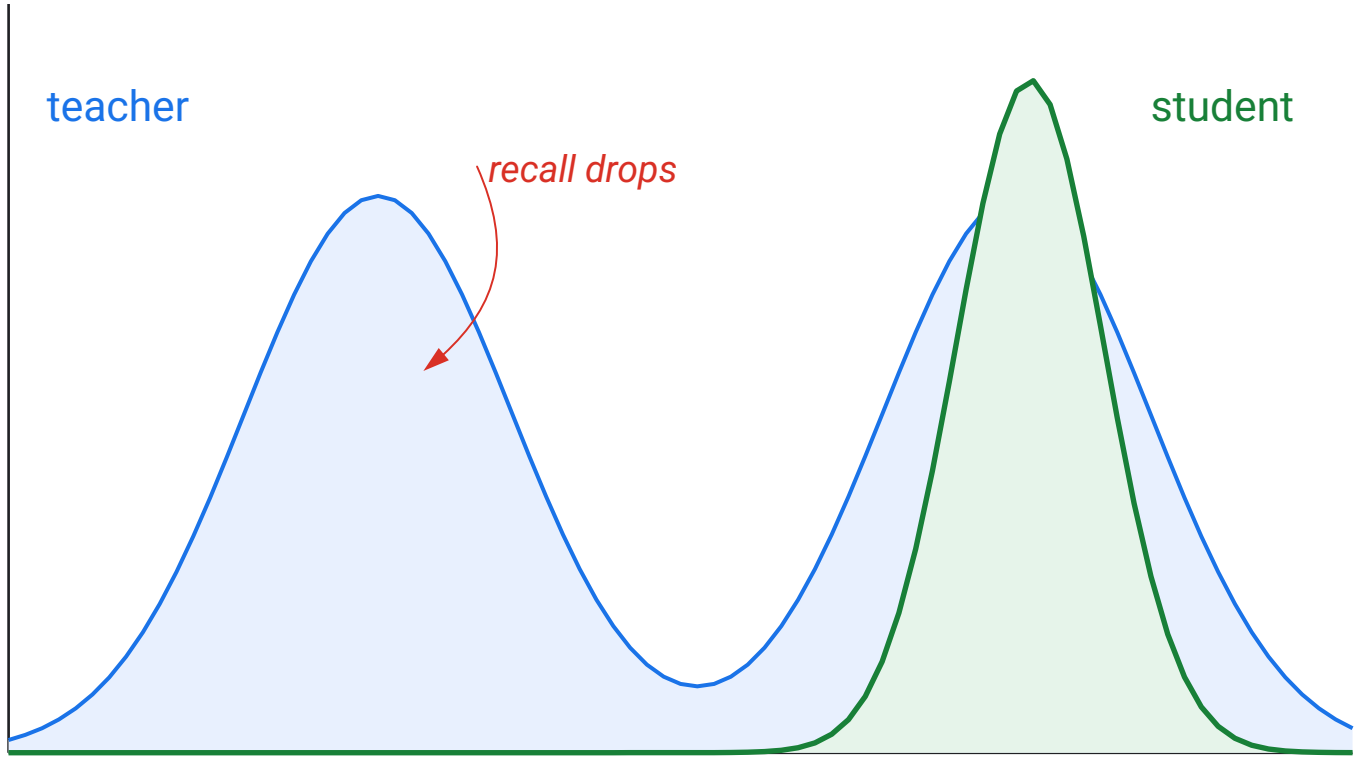
8 steps 1.24

teacher 1.42
at 1024 steps



8-step text-to-image

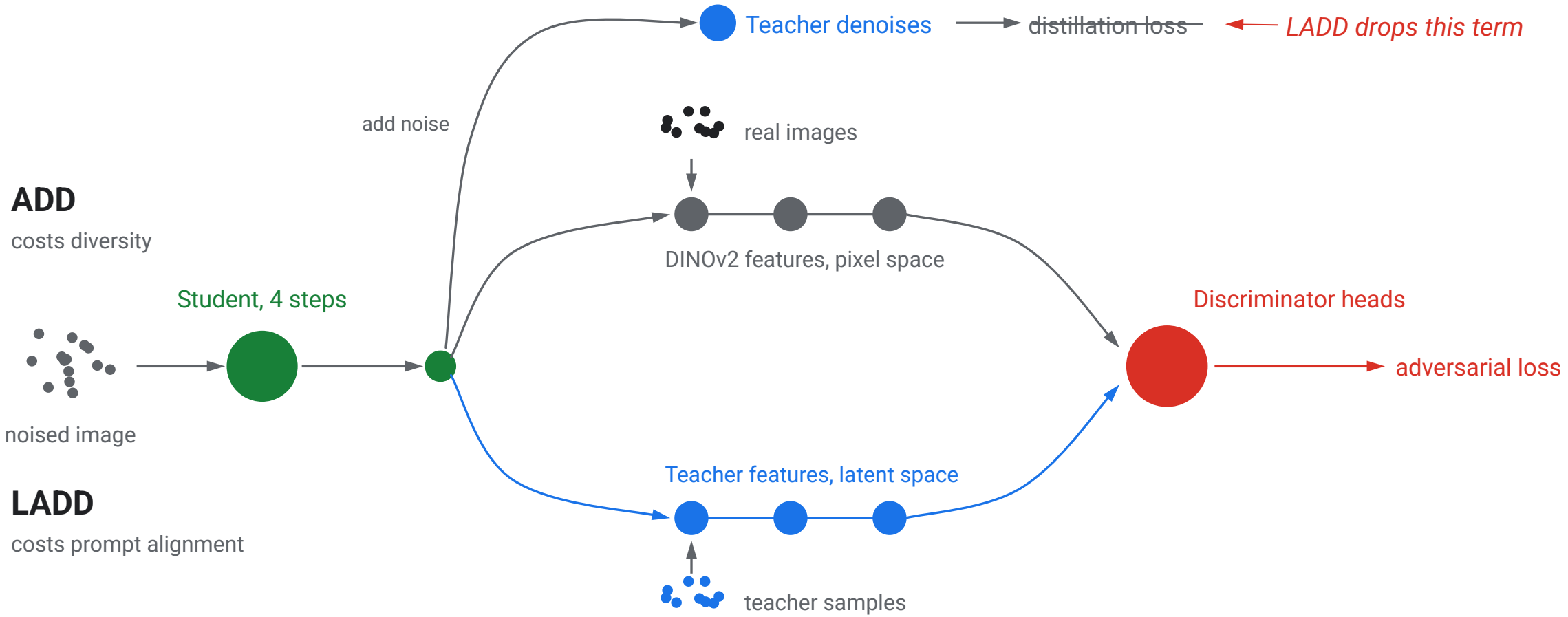
A student can beat its teacher's score



	teacher	student
FID, lower is better CIFAR-10	1.97	**1.923**
recall, higher is better ImageNet 64	0.67	**0.63**

SiD, one step, against its EDM teacher

LADD judges images in the teacher's features



[44] Sauer et al. 2023, arXiv:2311.17042 · [45] Sauer et al. 2024, arXiv:2403.12015

SDXL-Turbo at one, two and four steps

same seed



1 step | 2 steps | 4 steps



SDXL, 50 steps **SDXL-Turbo**, 4 steps

What few-step images still get wrong

garbled signs



pi-Flow, 8 NFE



FLUX Turbo, 8 NFE

wrong digits



teacher



sCM 4-step

burnt colour



LCM, w=4



w=12

one-step blur



256 steps



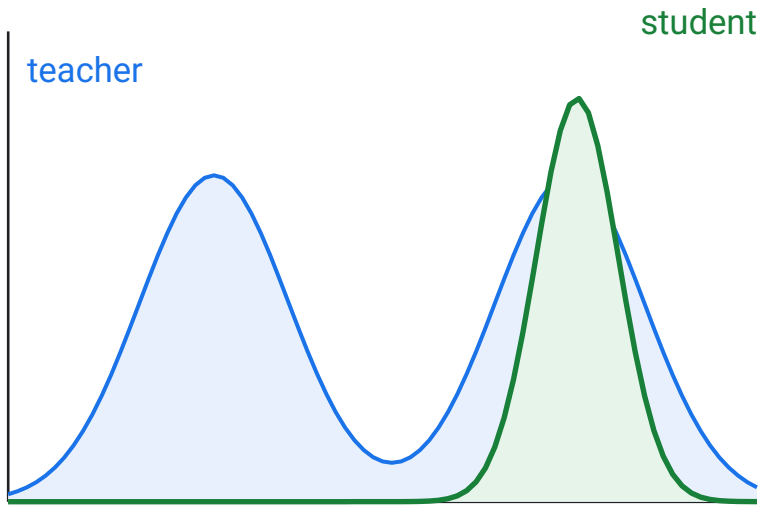
1 step

[74] Chen 2025, arXiv:2510.14974, Fig. 5 · [53] Zheng 2026, arXiv:2510.08431, Fig. 3 · [24] Luo 2023, arXiv:2310.04378, Fig. 5 · [6] Salimans 2022, arXiv:2202.00512, Fig. 3 · cropped

Why distilled models lose variety

the loss picks a peak

reverse KL



guidance sharpens the target

ImageNet 256, one-step **DMD**

guidance	FID	recall
none	2.13	0.64
1.5	5.20	0.48
4	15.70	0.11



[63] Jiang et al. 2025, arXiv:2511.13649 · [1] Yin et al. 2024, arXiv:2405.14867

New seed, same picture

SD3.5-M teacher, 60 NFE



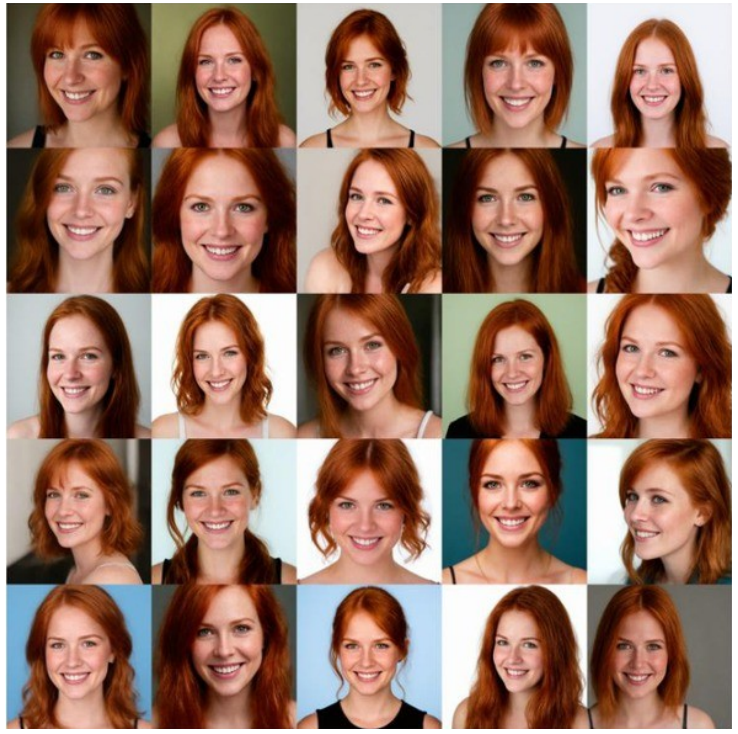
One prompt, 25 seeds per grid

DMD, 4 NFE



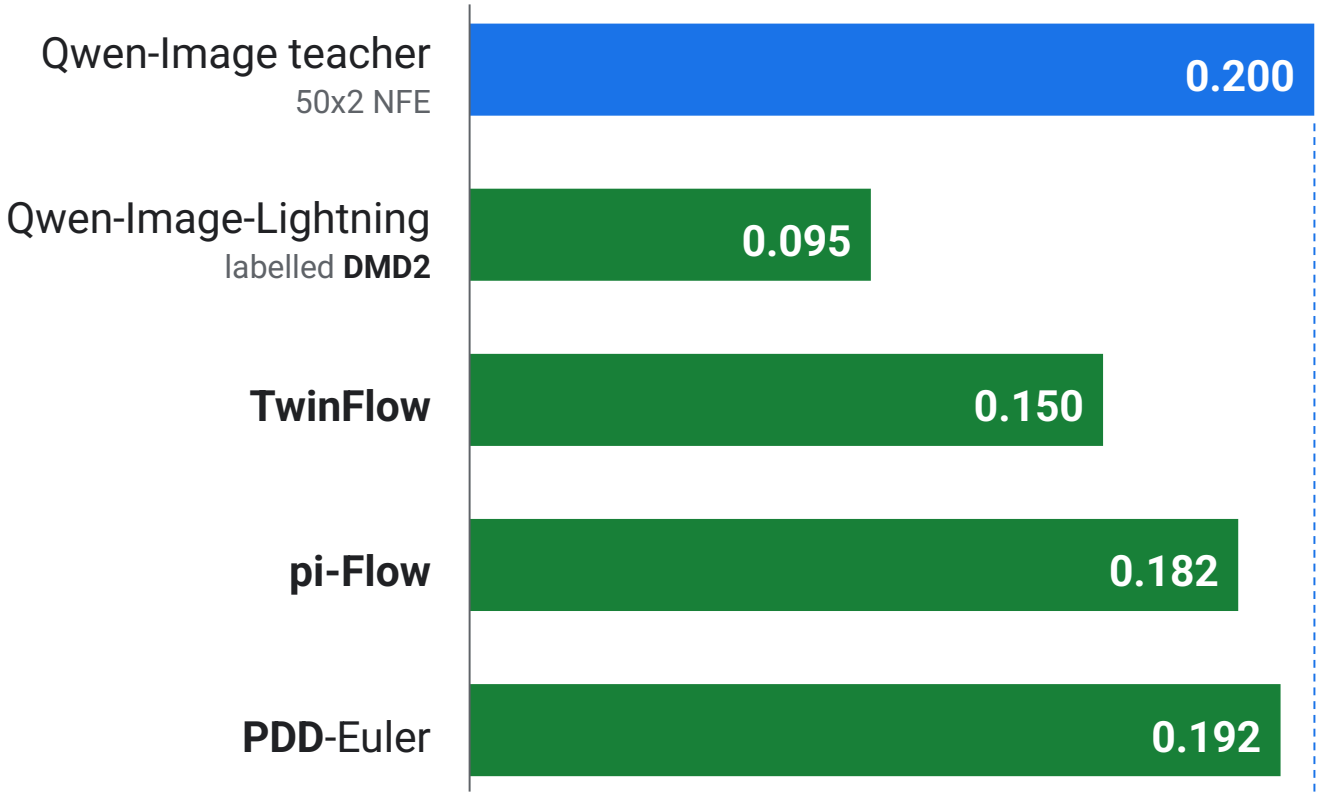
one face, 25 seeds

DP-DMD, 4 NFE



[80] Wu et al. 2026, arXiv:2602.03139, Fig. 1 · panels cropped and relabelled

Newer methods keep more of the variety



OneIG diversity, 4 NFE, one evaluator

Base model takes step 1

DreamSim 0.264 to 0.350, base 0.337

DP-DMD's step 1, SDXL

DINOv3 0.109 to 0.173, base 0.219

[73] Shaul 2026 · [74] Chen 2025 · [79] Cheng 2025 · [65] Gandikota 2025 · [80] Wu 2026

What the big labs ship for images

		steps	family
NVIDIA	**SANA-Sprint**	1 to 4	trajectory · discriminator
	Align Your Flow	4 to 8	trajectory
ByteDance Seed	**SDXL-Lightning**	1 to 8	discriminator
	Hyper-SD	1 to 8	trajectory
Nankai, Alibaba	**CDM**	4	distribution
Black Forest Labs	**FLUX.2 [klein]**	4	*the card never says how*

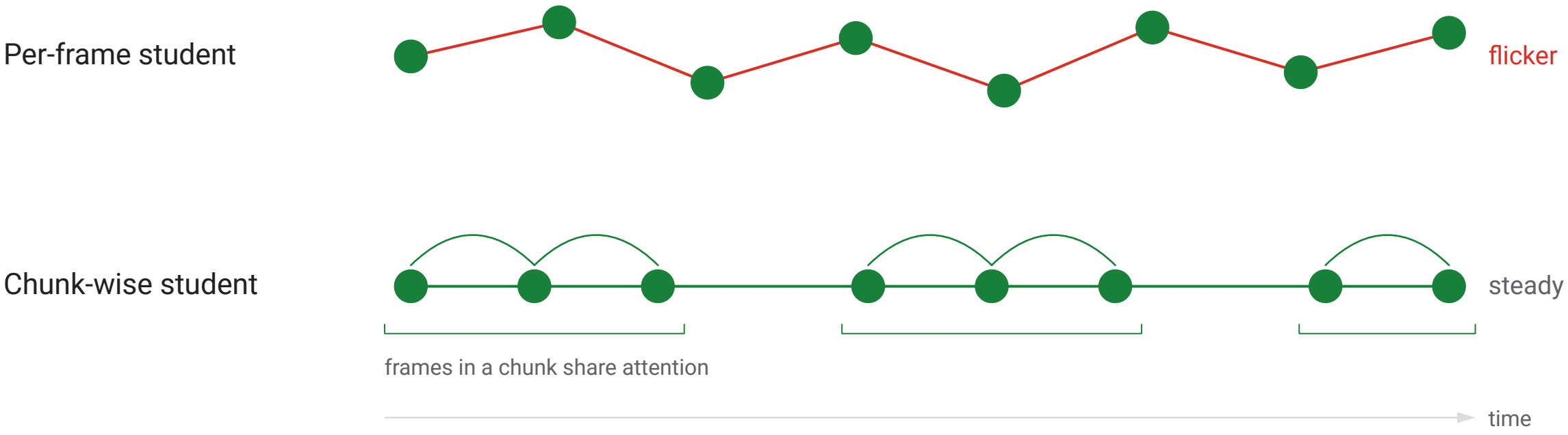
[16] Chen 2025 · [36] Sabour 2025 · [46] Lin 2024 · [26] Ren 2024 · [107] Liu 2026 · [108] Black Forest Labs 2026

PART 6

Distilling video models

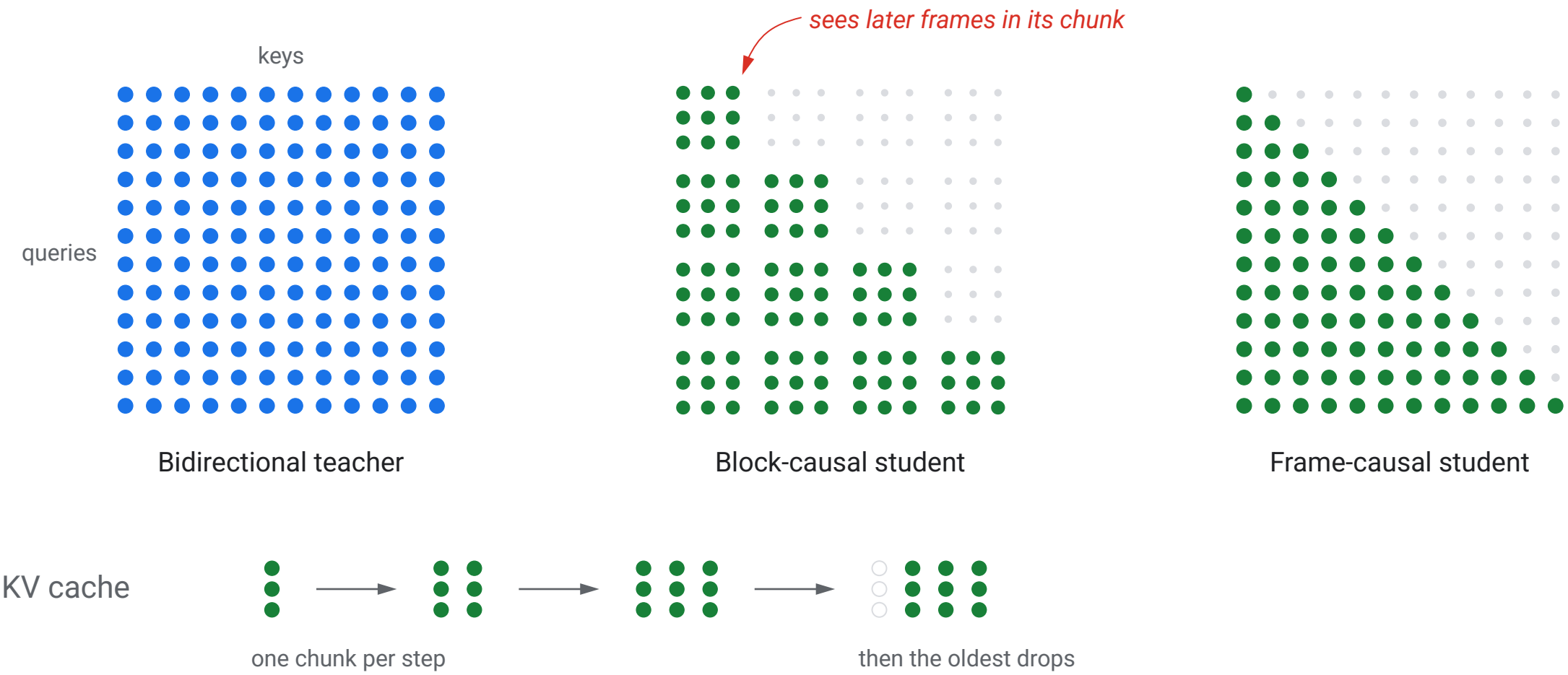
Why not run the image recipe on every frame?

Frame by frame, the video flickers



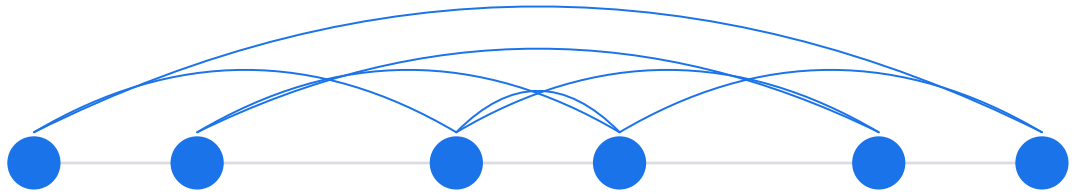
Wan2.1-1.3B takes 100 passes over 21 latent frames, 103 s per clip on one H100.

A causal mask lets the student stream

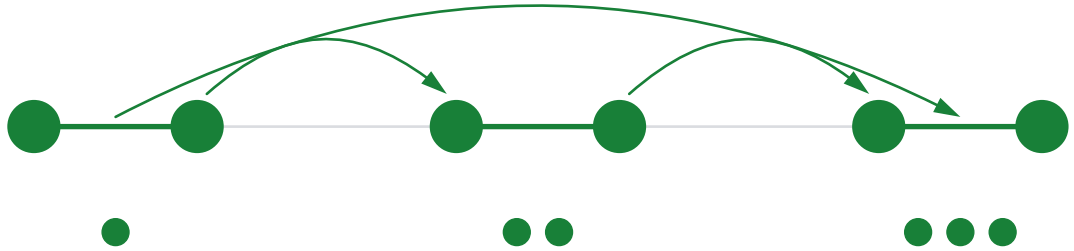


CausVid learns to stream from a bidirectional teacher

Bidirectional teacher



Causal student same weights, block-causal mask

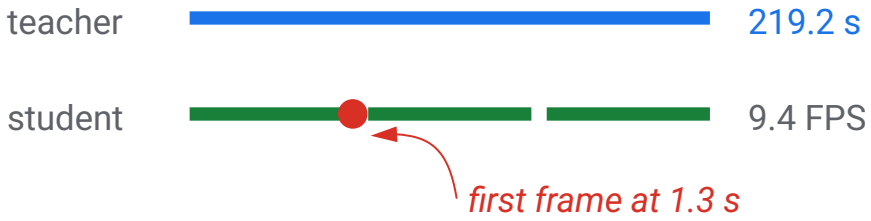


KV cache

Real score

Fake score

DMD loss



Four streaming steps next to fifty teacher steps

bidirectional teacher, 50 steps



CausVid, 4 steps



0 1.3 2.6 3.9 s

Same prompt; the student streams from 1.3 s.

rCM keeps Wan2.1 sharp down to two steps

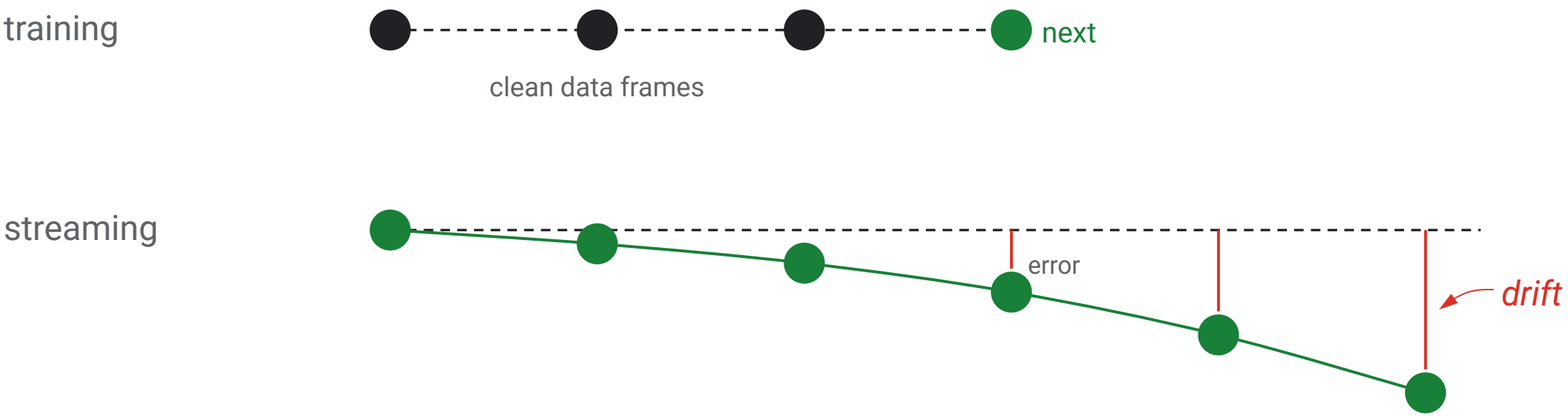


	VBench	FPS
1 step	82.65	32.3
2 steps	84.09	23.0
4 steps	84.43	14.6

One step blurs the street.

	VBench	FPS
teacher, 50 steps, CFG	83.02	0.72

Why small mistakes pile up in video



In training the student continues from clean past frames. When streaming it continues from its own, and small errors add up.

Exposure bias returns once the student streams

Off-policy context

Teacher forcing

ground truth

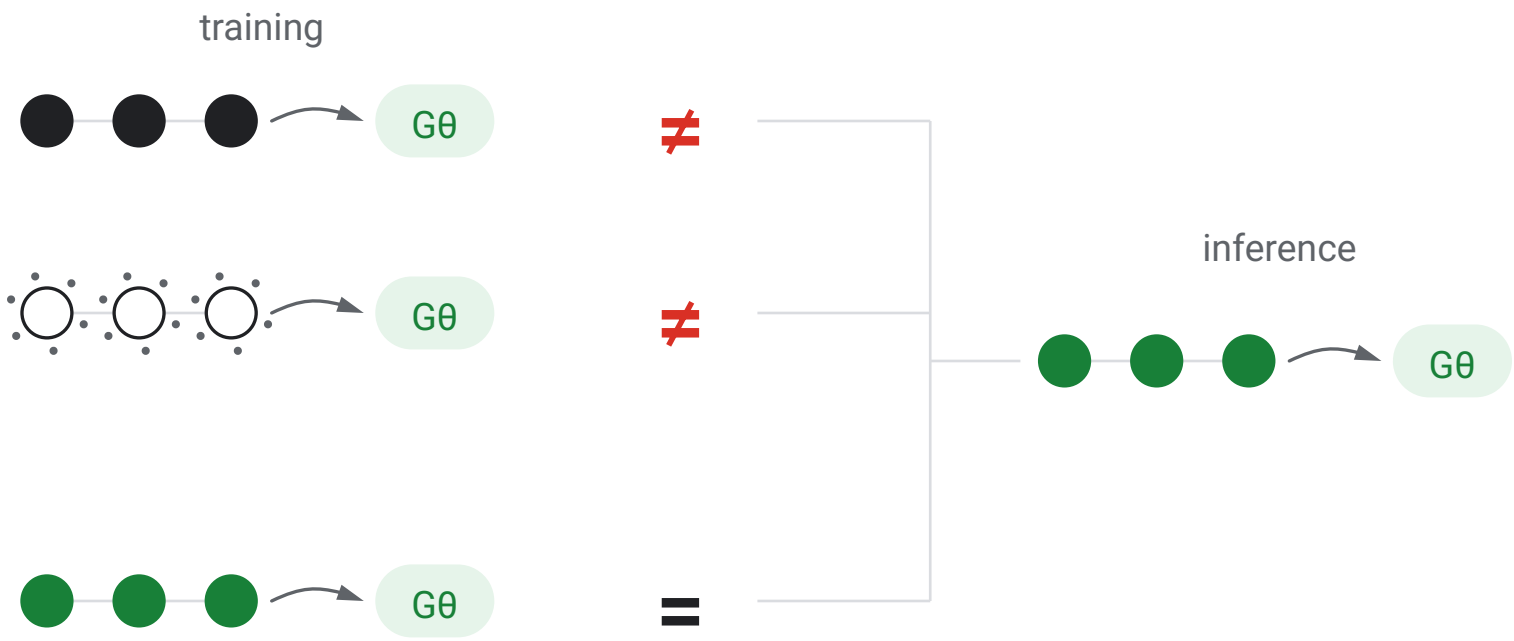
Diffusion forcing

noised ground truth

On-policy context

Self Forcing

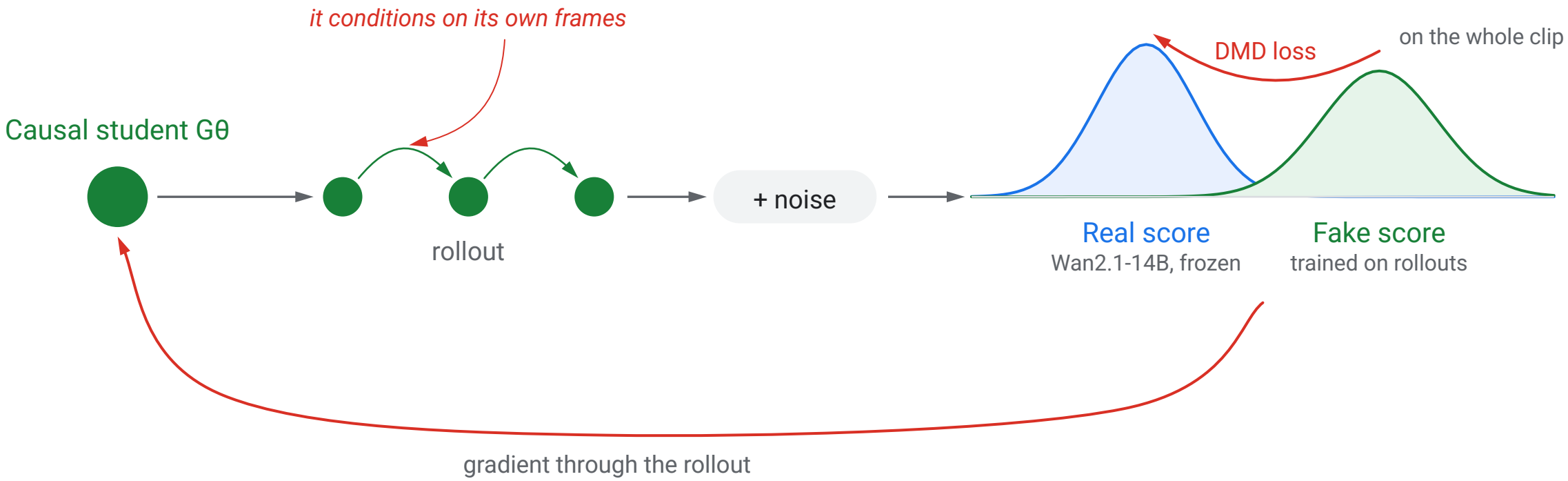
its own rollout



Same Wan2.1-1.3B base, VBench total: Self Forcing 84.31, **CausVid**'s recipe 81.20.

[19] Huang et al. 2025, arXiv:2506.08009 · [50] Yin et al. 2025, arXiv:2412.07772

Self Forcing trains on its own rollouts



On one H100, 17.0 FPS and a 0.69 s first frame. VBench 84.31.

CausVid oversaturates, Self Forcing does not

Wan2.1-1.3B teacher

CausVid, 4 steps

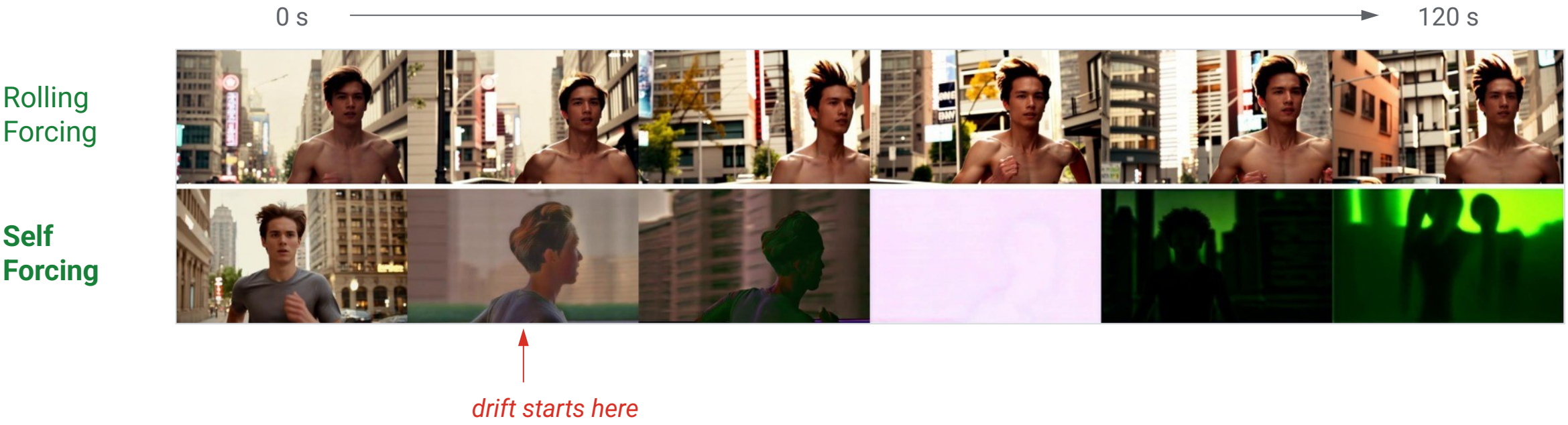
saturated, smoothed

Self Forcing, 4 steps



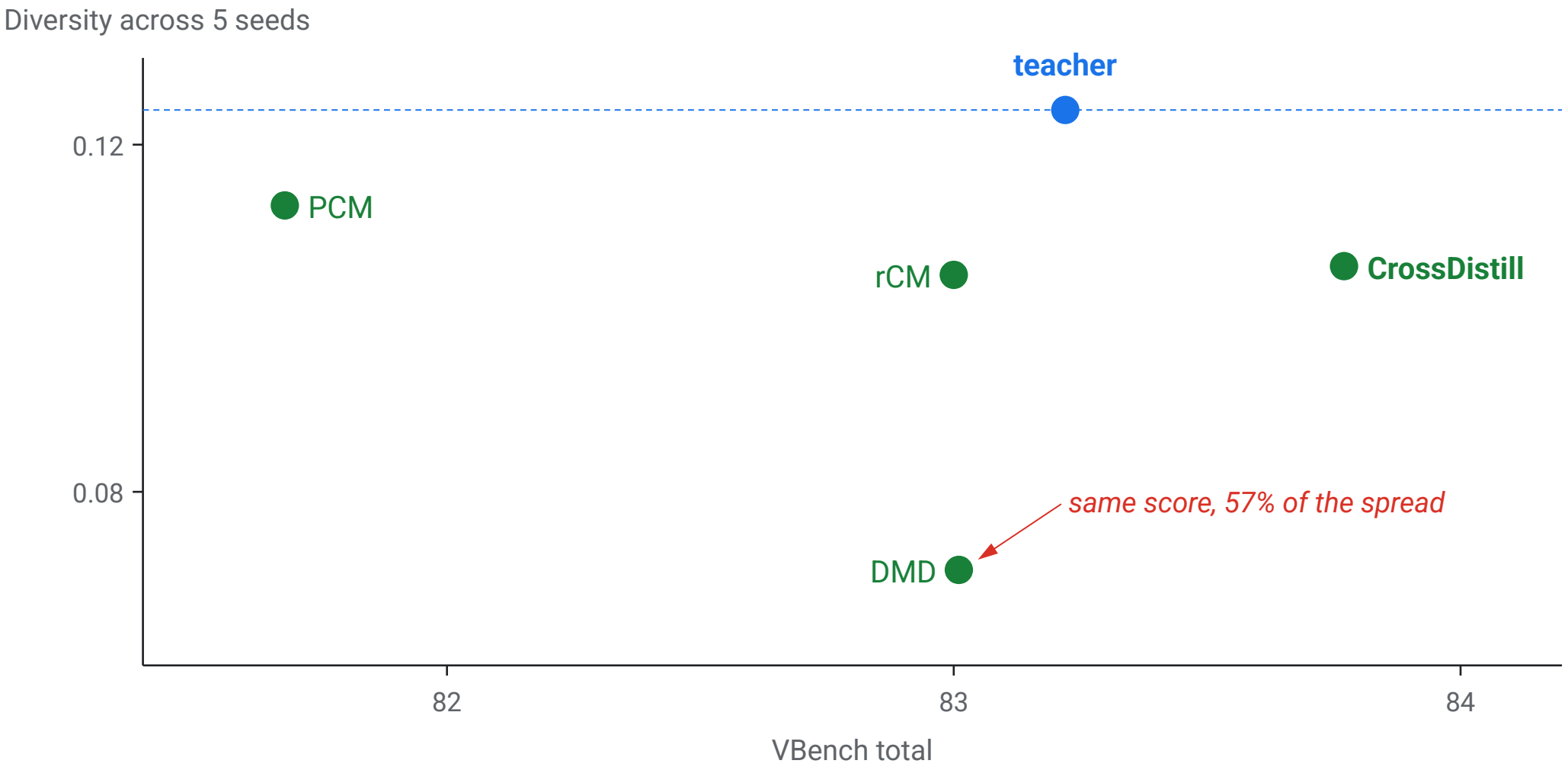
One prompt, four seconds

Self Forcing drifts after five seconds



Imaging drift over 30 s: **Self Forcing** 1.66, **Rolling Forcing** 0.01

Video students lose variety too



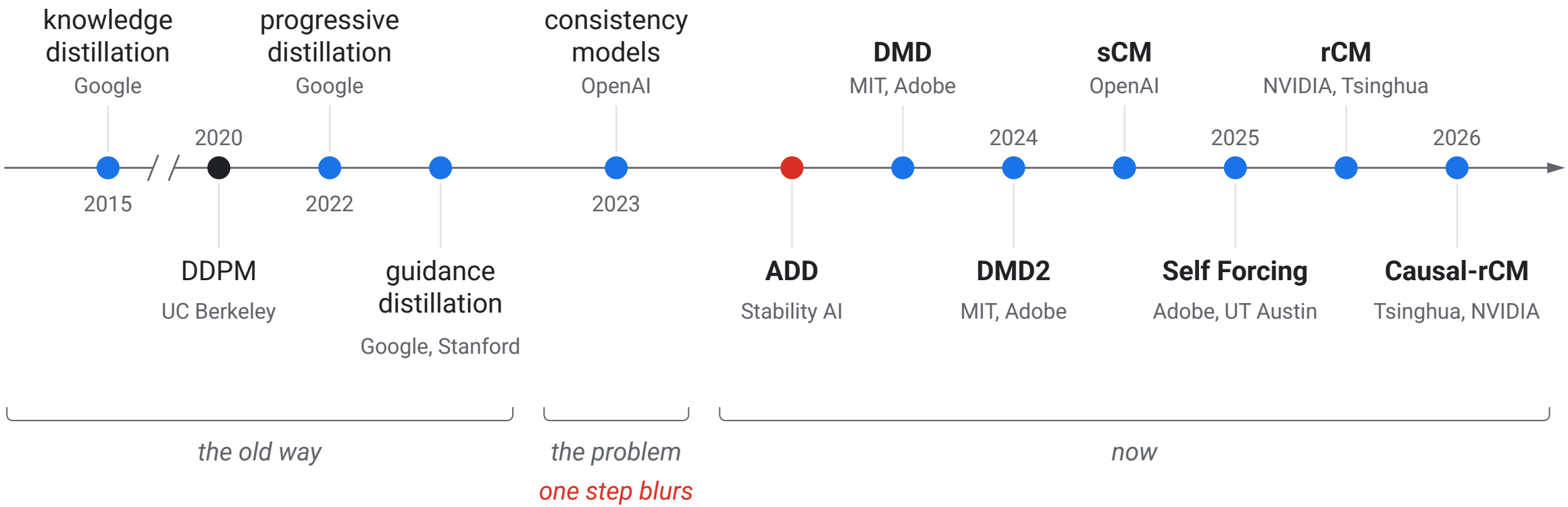
[83] CrossDistill 2026 · [53] Zheng 2026 · [84] PCM 2024 · [40] Yin 2024

What the big labs ship for video

		steps	family			
NVIDIA	**rCM**	1 to 4	trajectory	distribution		
	Cosmos-Predict2.5	4	trajectory	distribution		
	Causal-rCM	2 to 4	trajectory	distribution	streams	
ByteDance Seed	**APT**	1	discriminator			
	APT2	1	discriminator	streams	*one step per latent frame*	
others	**Self Forcing**	4	distribution	streams		
	Wan2.2-Lightning	4	distribution			

[53] Zheng 2026 · [113] NVIDIA 2025 · [81] Zheng 2026 · [47] Lin 2025 · [64] Lin 2025 · [19] Huang 2025 · [17] ModelTC 2025

Distillation from 2015 to 2026



PART 7

What still breaks

What do the scores not show?

Problems nobody has solved yet

Long videos drift

LongLive, **Self Forcing++**

Variety drops and is hard to measure

one DMD2, three LPIPS values

pi-Flow, **DP-DMD**

FID and GenEval disagree

SD3.5-Flash, DMD2

One-step video still needs a discriminator

APT, **APT2**

Latency numbers skip the VAE

SDXS, APT

[60] Yang 2025 · [1] Yin 2024 · [15] Bandyopadhyay 2025 · [47] Lin 2025 · [64] Lin 2025

Where I would start

images

1. SDXL teacher
2. **DMD2**
3. 5 fake-score updates
4. 4 steps
5. check variety

about 60 h on 64 A100s

video

1. Wan teacher
2. ODE init
3. **Self Forcing**
4. rolling cache
5. check drift

about 600 iterations on 64 H100s

Give the teacher the same wall-clock before you call it a win.

[1] Yin 2024 · [19] Huang 2025 · [49] Team Wan 2025

Using a distilled model takes a few lines

```
pipe = DiffusionPipeline.from_pretrained(
    "stabilityai/stable-diffusion-xl-base-1.0",
    torch_dtype=torch.float16, variant="fp16").to("cuda")
pipe.load_lora_weights(hf_hub_download(
    "tianweiy/DMD2", "dmd2_sdxl_4step_lora_fp16.safetensors"))
pipe.fuse_lora(lora_scale=1.0)
pipe.scheduler = LCMScheduler.from_config(pipe.scheduler.config)
image = pipe(prompt="a photo of a cat",
    num_inference_steps=4, guidance_scale=0,
    timesteps=[999, 749, 499, 249]).images[0]
```

LCM defaults differ from training

1. load teacher pipeline
2. swap in student weights
3. 4 steps, guidance off



[1] Yin 2024 · model card: huggingface.co/tianweiy/DMD2 · image CC BY 4.0

Which kind of distillation to pick

Family	Matches	Target from	Trained on	Weak spot
Trajectory PD, CD, LCM, reflow	a teacher ODE point	teacher	off-policy	few-step blur, teacher ceiling
Self-distillation CT, shortcut, MeanFlow	its own stop-gradient output	self	off-policy	from scratch, ImageNet only
Distribution DMD2, SiD, Self Forcing	teacher score minus fake score	teacher	on-policy	mode dropping, frozen guidance
Discriminator ADD, LADD	a discriminator's verdict	teacher	on-policy	collapse, lower diversity



[6] Salimans 2022 · [34] Geng 2025 · [1] Yin 2024 · [19] Huang 2025 · [45] Sauer 2024

References

[1] Yin 2024, arXiv:2405.14867
[2] Blau 2018, arXiv:1711.06077
[3] Robbins 1956, projecteuclid.org
[4] Efron 2011, doi.org
[5] Dieleman 2024, sander.ai
[6] Salimans 2022, arXiv:2202.00512
[7] Hinton 2015, arXiv:1503.02531
[8] Kim 2016, arXiv:1606.07947
[9] Gu 2024, arXiv:2306.08543
[10] Agarwal 2024, arXiv:2306.13649
[11] Lu 2025, thinkingmachines.ai
[12] Lightricks 2026, huggingface.co
[13] Black Forest Labs 2024, huggingface.co
[14] Stability AI 2024, huggingface.co
[15] Bandyopadhyay 2025, arXiv:2509.21318
[16] Chen 2025, arXiv:2503.09641
[17] ModelTC 2025, github.com
[18] Huang 2025, github.com
[19] Huang 2025, arXiv:2506.08009
[20] Meng 2023, arXiv:2210.03142
[21] Dieleman 2022, sander.ai
[22] Bradley 2024, arXiv:2408.09000
[23] Karras 2024, arXiv:2406.02507
[24] Luo 2023, arXiv:2310.04378
[25] Luo 2023, arXiv:2311.05556
[26] Ren 2024, arXiv:2404.13686
[27] Song 2023, arXiv:2303.01469
[28] Song 2024, arXiv:2310.14189
[29] Kim 2024, arXiv:2310.02279
[30] Lu 2025, arXiv:2410.11081
[31] Liu 2022, arXiv:2209.03003
[32] Liu 2024, arXiv:2309.06380
[33] Frans 2025, arXiv:2410.12557
[34] Geng 2025, arXiv:2505.13447
[35] Boffi 2024, arXiv:2406.07507
[36] Sabour 2025, arXiv:2506.14603
[37] Zhou 2025, arXiv:2503.07565
[38] Wang 2023, arXiv:2305.16213
[39] Luo 2023, arXiv:2305.18455
[40] Yin 2024, arXiv:2311.18828
[41] Zhou 2024, arXiv:2404.04057
[42] Zhou 2025, arXiv:2406.01561
[43] Xie 2024, arXiv:2405.16852
[44] Sauer 2023, arXiv:2311.17042
[45] Sauer 2024, arXiv:2403.12015
[46] Lin 2024, arXiv:2402.13929
[47] Lin 2025, arXiv:2501.08316
[48] Xu 2023, arXiv:2311.09257
[49] Team Wan 2025, arXiv:2503.20314
[50] Yin 2025, arXiv:2412.07772
[51] Li 2024, arXiv:2405.18750
[52] Li 2025, arXiv:2410.05677
[53] Zheng 2026, arXiv:2510.08431
[54] HaCohen 2025, arXiv:2501.00103
[55] Kynkäänniemi 2019, arXiv:1904.06991
[56] Naeem 2020, arXiv:2002.09797
[57] Ghosh 2023, arXiv:2310.11513
[58] Song 2024, arXiv:2403.16627
[59] Poole 2022, arXiv:2209.14988
[60] Yang 2025, arXiv:2509.22622
[61] Liu 2025, arXiv:2509.25161
[62] Cui 2025, arXiv:2510.02283
[63] Jiang 2025, arXiv:2511.13649
[64] Lin 2025, arXiv:2506.09350
[65] Gandikota 2025, arXiv:2503.10637
[66] Huang 2024, arXiv:2311.17982
[67] Zhang 2025, arXiv:2510.20771
[68] Lee 2025, arXiv:2510.24474
[69] Geng 2025, arXiv:2512.02012
[70] Lu 2025, arXiv:2507.18569
[71] Ge 2025, arXiv:2506.00523
[72] Chang 2025, arXiv:2506.07977
[73] Shaul 2026, arXiv:2607.26004
[74] Chen 2025, arXiv:2510.14974
[75] Salimans 2024, arXiv:2406.04103
[76] Zhou 2024, arXiv:2410.14919
[77] Xu 2025, arXiv:2502.15681
[78] Wang 2025, arXiv:2505.20755
[79] Cheng 2025, arXiv:2512.05150
[80] Wu 2026, arXiv:2602.03139
[81] Zheng 2026, arXiv:2606.25473
[82] Zhu 2026, arXiv:2602.02214
[83] Liu 2026, arXiv:2609.14725
[84] Wang 2024, arXiv:2405.18407
[85] Duan 2026, arXiv:2609.11265
[86] Fan 2025, arXiv:2510.27684
[107] Liu 2026, arXiv:2605.06376
[108] Black Forest Labs 2026, huggingface.co
[113] NVIDIA 2025, arXiv:2511.00062

Example images: [1], [44] CC BY 4.0; [19], [50] CC BY-NC-SA 4.0; [32] MIT; [53] Apache-2.0; web figures as cited. Cropped and combined.

Thanks

shreshthsaini.github.io/blogs/diffusion-distillation.html

Questions