COMP 690 · DEEP GENERATIVE MODELS · READING: DIFFUSION

Progressive Distillation for Fast Sampling of Diffusion Models

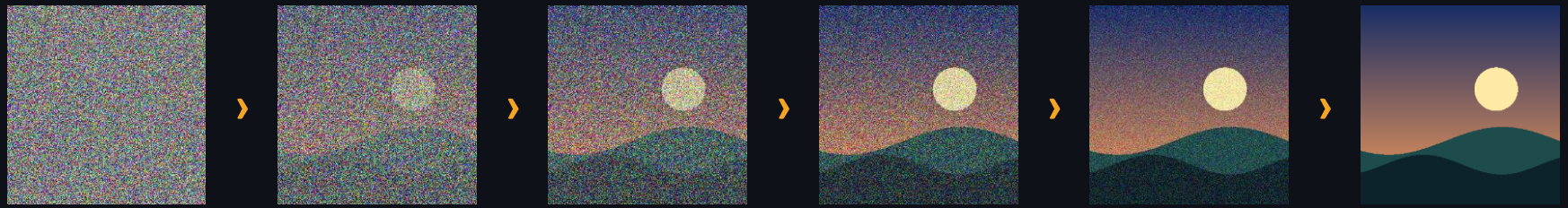
1024 → 4

sampling steps, little quality loss

Tim Salimans & Jonathan Ho

Google Research, Brain Team · ICLR 2022

Arushi Sahai · Tue, Oct 13, 2026



Illustrative: a synthetic image at decreasing noise levels (not model output)

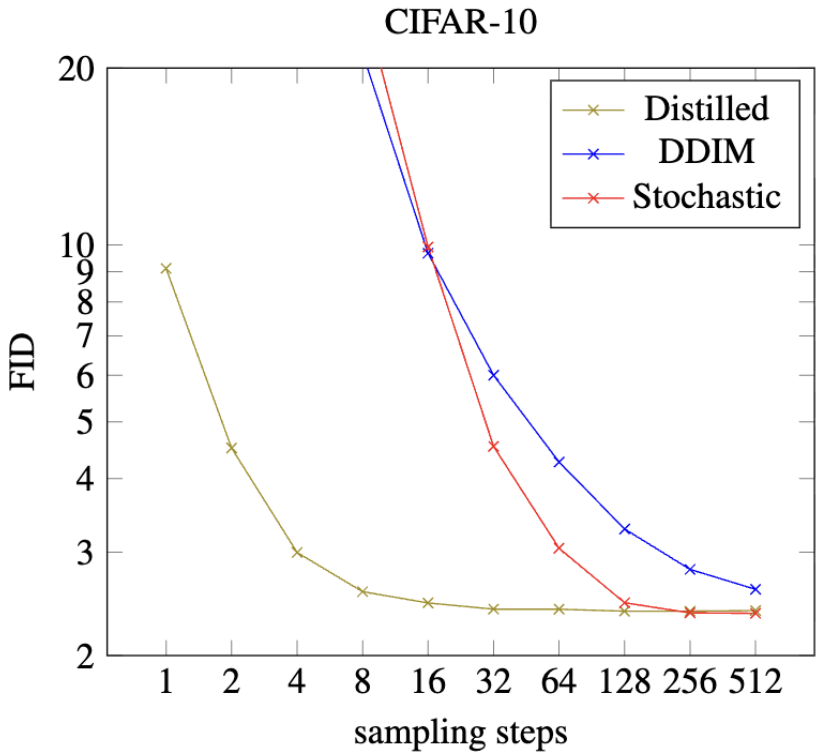
Diffusion wins on quality, not speed

	VAE	GAN	Diffusion (DDPM)
Sample quality	**Blurry averages**	Sharp	Sharp; beats GANs on FID
Mode coverage	Good	**Mode collapse**	Good
Training	Stable (ELBO)	**Unstable, adversarial**	Stable (simple L2 loss)
Sampling cost	1 network pass	1 network pass	**100s to 1000s of passes**

Diffusion avoids VAE blur and GAN instability. Its remaining weakness is sampling speed, the gap this paper closes.

Source: Salimans & Ho (2022), Sec. 1; COMP 690 Lec 13. Red text marks a weakness.

Fewer steps alone costs quality



Use fewer steps
DDIM and stochastic samplers degrade below ~128 steps

Re-tune the sampler
Works only with strong conditioning, e.g. super-resolution

Distill to one step
Needs full teacher runs; cost grows with teacher steps

Salimans & Ho (2022), Fig. 4.

Same model, different notation

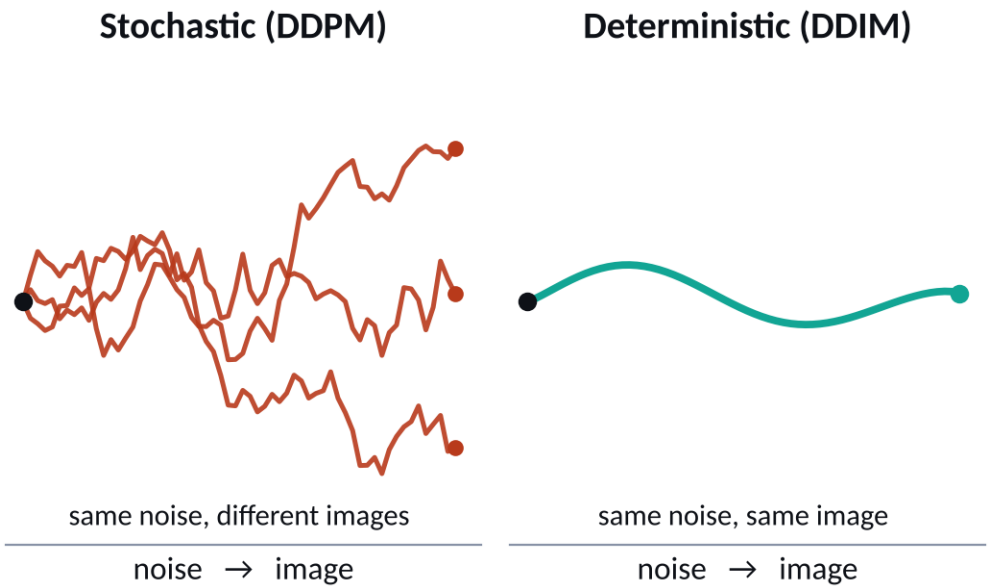
Quantity	Lecture 13 (DDPM)	This paper
Time	$t = 1, \ldots, T$ (discrete)	$t \in [0, 1]$ (continuous)
Noisy sample	$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{(1-\bar{\alpha}_t)}\, \varepsilon$	$z_t = \alpha_t\, x + \sigma_t\, \varepsilon$
Schedule	β_t, $\bar{\alpha}_t$	α_t, σ_t, log-SNR λ_t
Network predicts	$\varepsilon_\theta(x_t, t)$	$\hat{x}_\theta(z_t)$, or $\hat{v}$
Loss	$\lVert\varepsilon - \varepsilon_\theta\rVert^2$, weight 1	$w(\lambda_t)\, \lVert x - \hat{x}_\theta(z_t)\rVert^2$
Sampler	Ancestral (stochastic)	DDIM (deterministic)

Mapping: $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{(1-\bar{\alpha}_t)}$. The lecture's ε-loss is this loss with w = SNR (Eq. 9).

Salimans & Ho (2022), Sec. 2 and Eq. 9; COMP 690 Lec 13.

DDIM: a deterministic sampler

Denoising Diffusion Implicit Model



Same noise, same image

No fresh noise per step, so the noise-to-image map is fixed

An ODE integrator

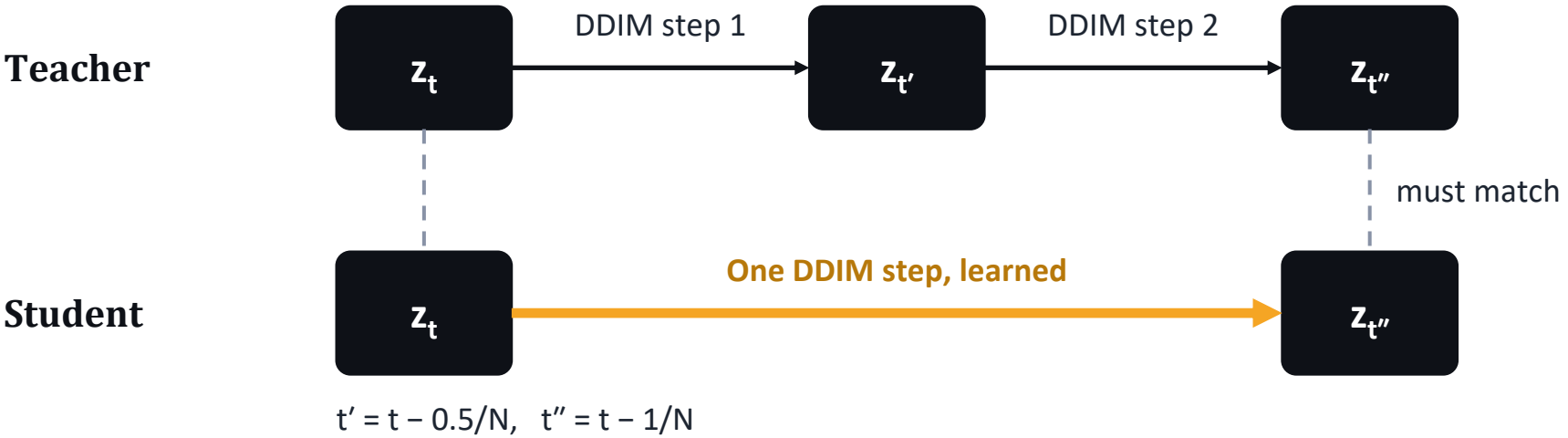
DDIM follows the probability-flow ODE to first order (App. B)

DDIM update (Eq. 7)

$$z_s = \alpha_s \hat{x} + \sigma_s \left(z_t - \alpha_t \hat{x}\right) / \sigma_t$$

Illustrative paths. Salimans & Ho (2022), Sec. 2 and App. B. DDIM is covered in full on Oct 29.

Distill two teacher steps into one



1. Teacher takes two DDIM steps
2. Student must reach the same point in one
3. Student becomes teacher; N → N/2

After Salimans & Ho (2022), Fig. 1 and Sec. 3. Diagram redrawn.

Training loop barely changes

Algorithm 1 Standard diffusion training

Require: Model $\hat{\mathbf{x}}_\theta(\mathbf{z}_t)$ to be trained
Require: Data set $\mathcal{D}$
Require: Loss weight function $w()$

while not converged **do**
- $\mathbf{x} \sim \mathcal{D}$ ▷ Sample data
- $t \sim U[0, 1]$ ▷ Sample time
- $\epsilon \sim N(0, I)$ ▷ Sample noise
- $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$ ▷ Add noise to data
- $\tilde{\mathbf{x}} = \mathbf{x}$ ▷ Clean data is target for $\hat{\mathbf{x}}$
- $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ ▷ log-SNR
- $L_\theta = w(\lambda_t)\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t)\|_2^2$ ▷ Loss
- $\theta \leftarrow \theta - \gamma \nabla_\theta L_\theta$ ▷ Optimization

end while

Algorithm 2 Progressive distillation

Require: Trained teacher model $\hat{\mathbf{x}}_\eta(\mathbf{z}_t)$
Require: Data set $\mathcal{D}$
Require: Loss weight function $w()$
Require: Student sampling steps N

for K iterations **do**
- $\theta \leftarrow \eta$ ▷ Init student from teacher
- **while** not converged **do**
  - $\mathbf{x} \sim \mathcal{D}$
  - $t = i/N, \; i \sim Cat[1, 2, \ldots, N]$
  - $\epsilon \sim N(0, I)$
  - $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$
  - `# 2 steps of DDIM with teacher`
  - $t' = t - 0.5/N, \quad t'' = t - 1/N$
  - $\mathbf{z}_{t'} = \alpha_{t'} \hat{\mathbf{x}}_\eta(\mathbf{z}_t) + \frac{\sigma_{t'}}{\sigma_t}(\mathbf{z}_t - \alpha_t \hat{\mathbf{x}}_\eta(\mathbf{z}_t))$
  - $\mathbf{z}_{t''} = \alpha_{t''} \hat{\mathbf{x}}_\eta(\mathbf{z}_{t'}) + \frac{\sigma_{t''}}{\sigma_{t'}}(\mathbf{z}_{t'} - \alpha_{t'} \hat{\mathbf{x}}_\eta(\mathbf{z}_{t'}))$
  - $\tilde{\mathbf{x}} = \frac{\mathbf{z}_{t''} - (\sigma_{t''}/\sigma_t)\mathbf{z}_t}{\alpha_{t''} - (\sigma_{t''}/\sigma_t)\alpha_t}$ ▷ Teacher $\hat{\mathbf{x}}$ target
  - $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$
  - $L_\theta = w(\lambda_t)\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t)\|_2^2$
  - $\theta \leftarrow \theta - \gamma \nabla_\theta L_\theta$
- **end while**
- $\eta \leftarrow \theta$ ▷ Student becomes next teacher
- $N \leftarrow N/2$ ▷ Halve number of sampling steps

end for

Only the target changes

$\tilde{x}$ is what the student must predict so one step matches two teacher steps

A sharper target

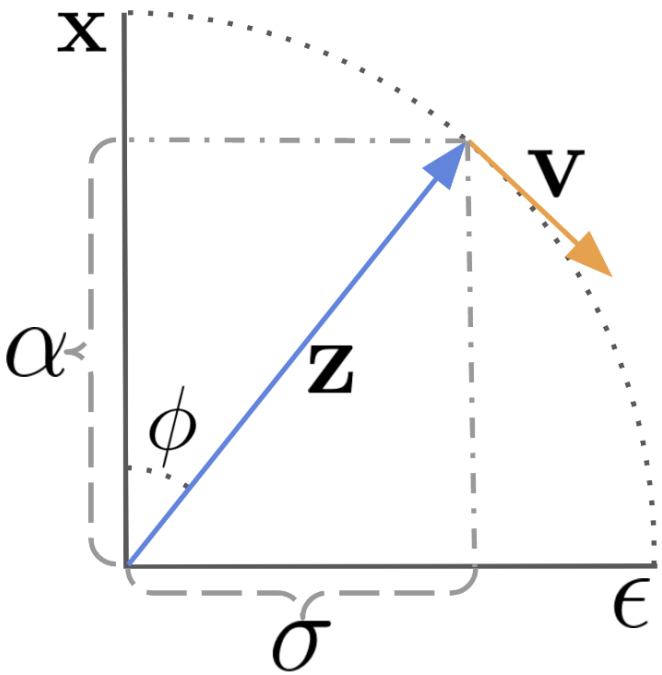
$\tilde{x}$ is fixed by z_t. The data x is not, so the old target is a blurry average

Each round halves N: 8192 (CIFAR-10) or 1024 (others) down to 1. 50k updates per round.



Salimans & Ho (2022), Sec. 3, Alg. 1 and 2, App. G. 100k updates for the 2- and 1-step rounds.

ε-prediction breaks at low SNR



The problem

$\hat{x} = (z_t - \sigma_t \hat{\epsilon})/\alpha_t$ divides by $\alpha_t \to 0$ at low SNR. At one step the input is pure noise, so $\hat{\epsilon}$ says nothing about x.

The fix: predict velocity v

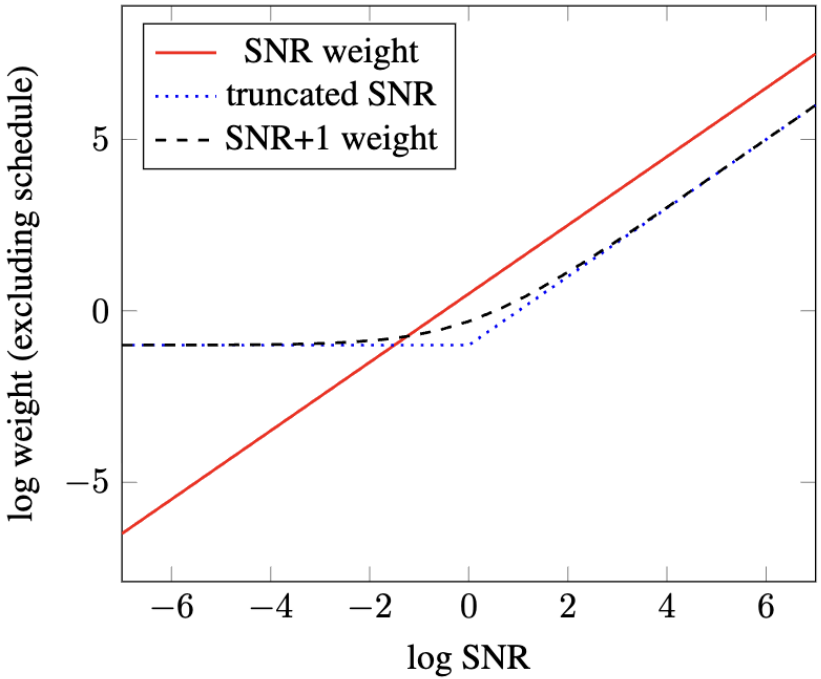
$v = \alpha_t \epsilon - \sigma_t x$, so $\hat{x} = \alpha_t z_t - \sigma_t \hat{v}$. Stable at every SNR.

Also stable

Predict x, or predict (x, ε) with two outputs. With v, the DDIM step size does not depend on SNR (App. D).

Salimans & Ho (2022), Sec. 4, Fig. 5 and App. D. Here ϕ = arctan(σ_t/α_t).

Loss weights must stay positive



CIFAR-10 FID, stochastic sampler

Output	SNR	Truncated	SNR+1
ε	2.59	**diverges**	2.56
x	2.65	2.53	2.56
v	2.65	2.45	2.49
(x, ε)	2.54	2.47	2.52

All stable choices land within about 0.2 FID. The ε output with truncated SNR diverges.

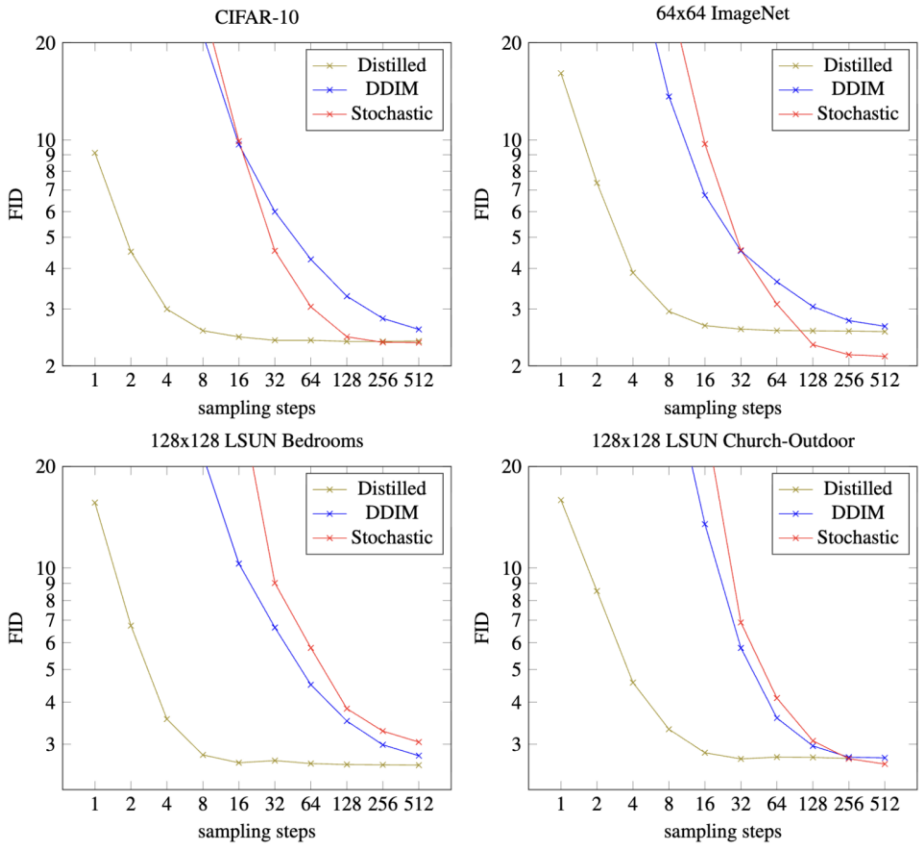
Salimans & Ho (2022), Sec. 4, Fig. 2, and Table 1.

Near-teacher quality at 4 to 8 steps

Setup

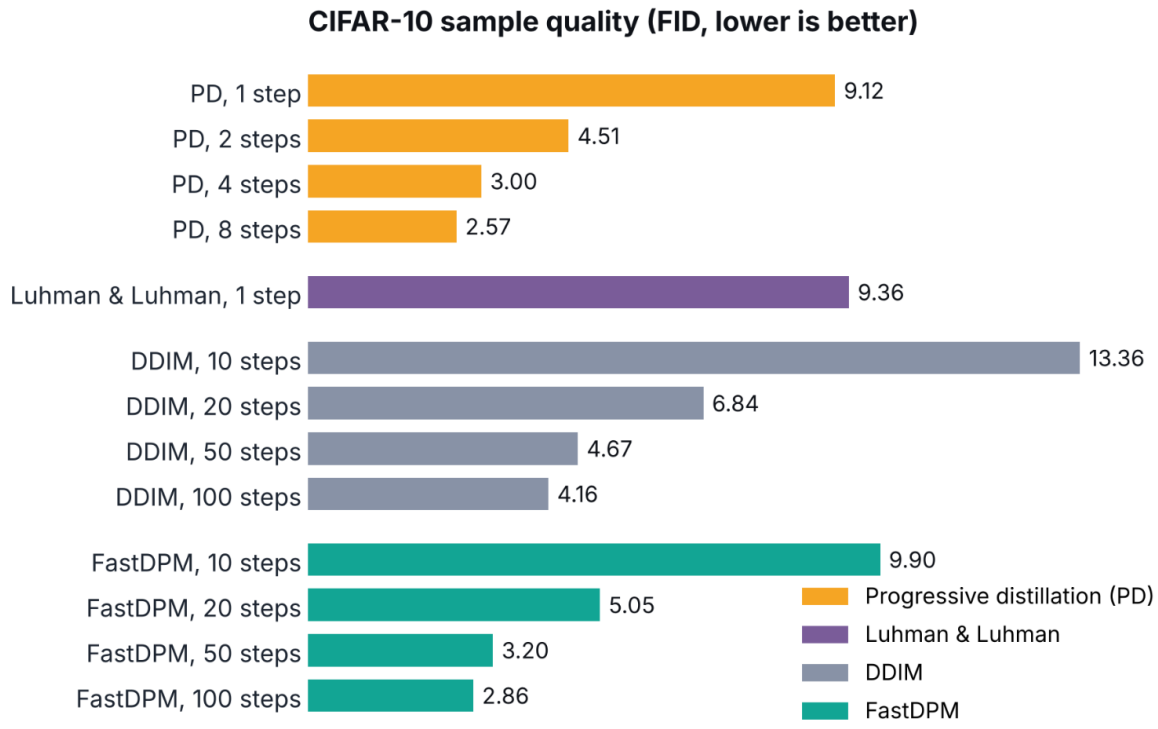
- CIFAR-10: 8192-step teacher
- ImageNet 64×64, LSUN bedroom and church 128×128: 1024-step teachers
- Same U-Net, 50k updates per halving
- Baselines: DDIM and a tuned stochastic sampler

Quality degrades at 2 and 1 steps. DDIM and stochastic samplers degrade sharply below 128 steps.



Salimans & Ho (2022), Sec. 5.2 and Fig. 4. FID: lower is better.

4 steps beat 100-step DDIM



3.00

PD, 4 steps

4.16

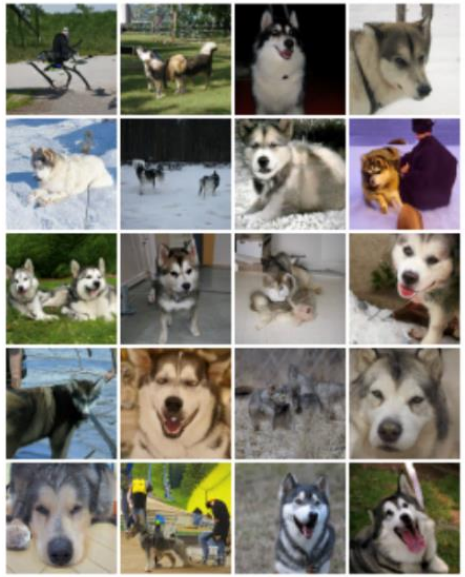
DDIM, 100 steps

9.12

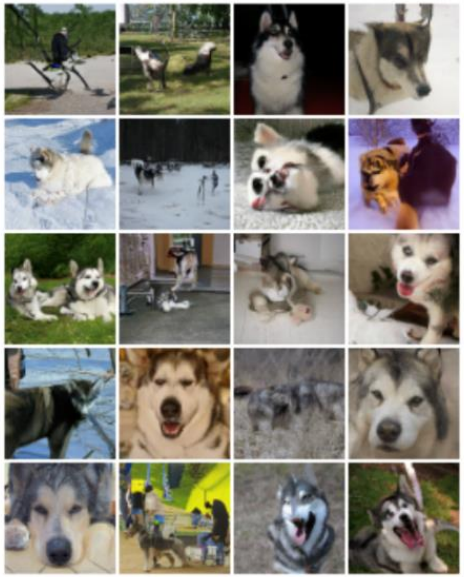
PD, 1 step (Luhman: 9.36)

Source: Salimans & Ho (2022), Table 2. Step counts are network evaluations per sample.

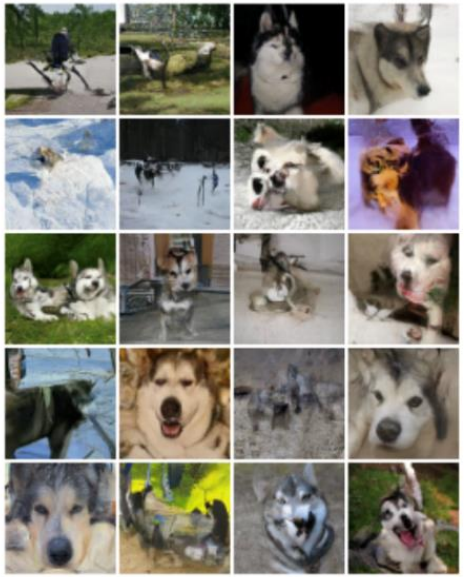
Same noise, same image, fewer steps



(a) 256 sampling steps



(b) 4 sampling steps



(c) 1 sampling step

The noise-to-image map is preserved; images lose some detail at 1 step.

Salimans & Ho (2022), Fig. 3: 64×64 ImageNet, class 'malamute', fixed seed

Big speedup, with real limits

Strengths

- Small change: same loss and network, new target
- Cost grows with log of teacher steps, not linearly
- Keeps the noise-to-image map and sample diversity
- About 1 to 5 days on TPUv4; code is open source

Weaknesses

- Needs deterministic DDIM; stochastic sampling loses fidelity
- Quality drops at 1 to 2 steps (FID 9.12 vs 3.00 at 4 steps)
- Student works only on its fixed step grid
- Tested on images up to 128×128; one seed on large sets

Salimans & Ho (2022), Secs. 5-7, App. E and G.

Where this goes next

1

Smaller students

The student matches the teacher's size here. Shrinking it is future work the authors name.

2

Other data types

The authors point to audio. Does halving work as well on other modalities?

3

Guidance and latent models

Can a student absorb guidance? See today's classifier-free guidance and latent diffusion papers.

4

Other routes to few steps

Straighter paths (flow matching, Oct 20) and consistency models (Oct 29).

Salimans & Ho (2022), Sec. 7; course schedule.

Takeaways

1. Sampling is an ODE; distillation learns to integrate it in fewer steps.

2. v-prediction and positive loss weights keep few-step training stable.

3. Halving N repeatedly reaches 4 steps (CIFAR-10 FID 3.0) for about the compute of the original training.

Salimans & Ho, ICLR 2022