DMD2 – Distribution Matching Distillation 2

Yu Chung (Paul) Lee, Arman Lotfalikhani

1 April 2026

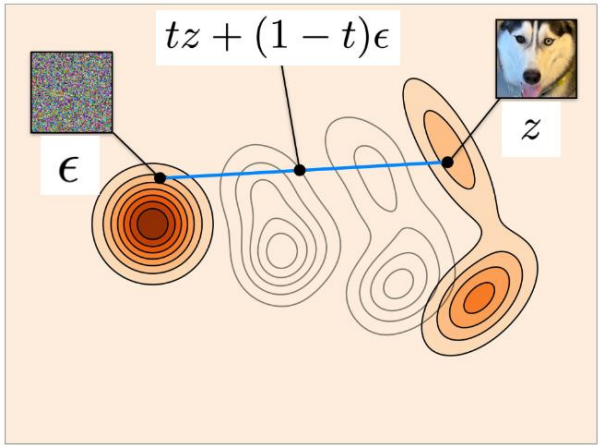
Improved Distribution Matching Distillation for Fast Image Synthesis

Tianwei Yin 1, Michaël Gharbi 2, Taesung Park 2, Richard Zhang 2,
Eli Shechtman 2, Frédo Durand 1, William T. Freeman 1

1 Massachusetts Institute of Technology, 2 Adobe Research
NeurIPS 2024 (Oral)

Motivation

- Diffusion models have been successful is visual generative tasks, while suffering from slow inference
- The iterative denoising procedure in diffusion models is the main bottleneck
- Idea: To use knowledge distillation methods to sample from the distribution in one or few steps generator



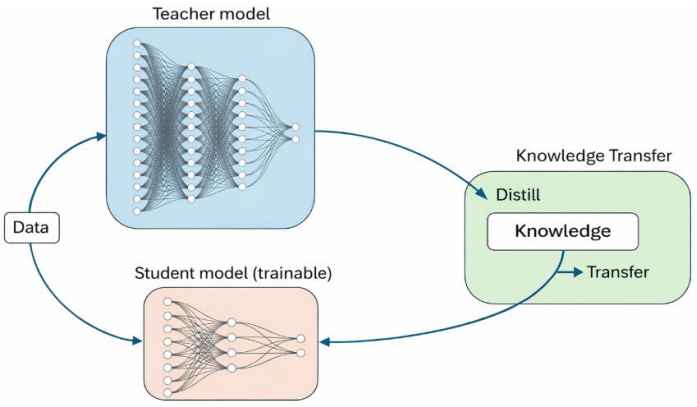
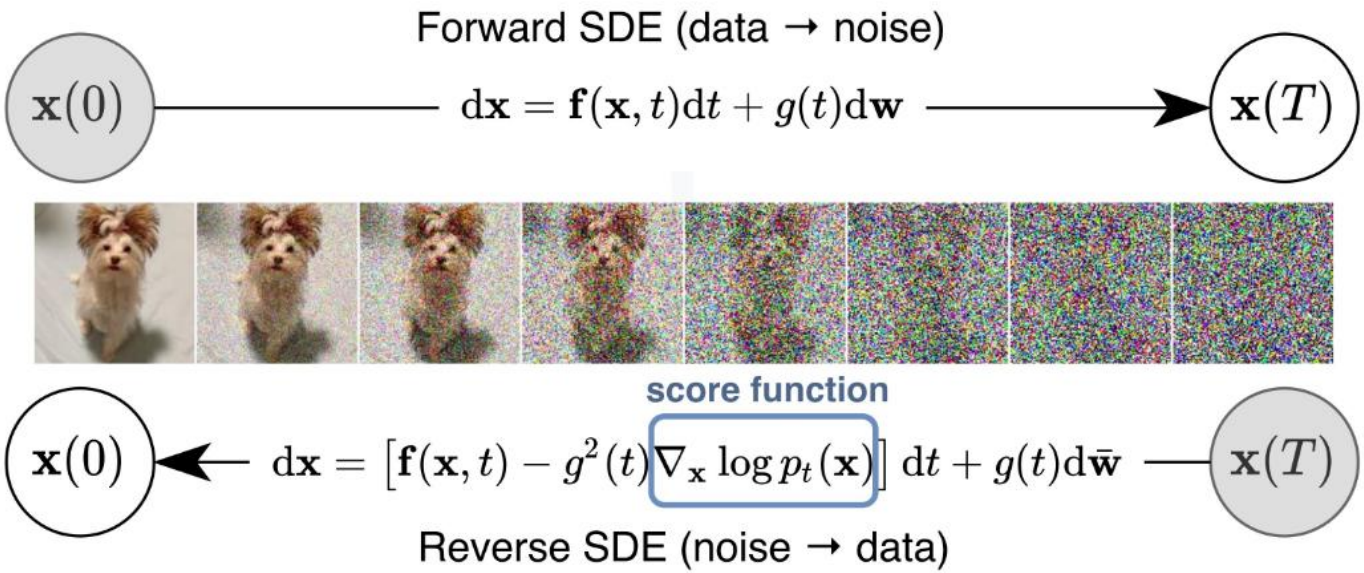


Figure credit:Yaron Lipman

Score-based Diffusion Model

Multiple steps denoising: 20~50 steps for reasonable generation



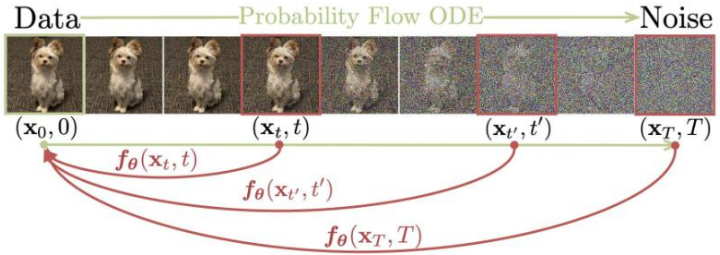
Probabilistic Machine Learning. Kevin Murphy (2023).

Trajectory Preserving Diffusion Distillation

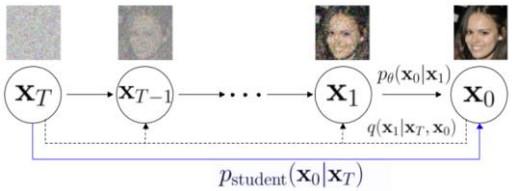
Learning ODE trajectories endpoint is viable but limited by one-step generator capability

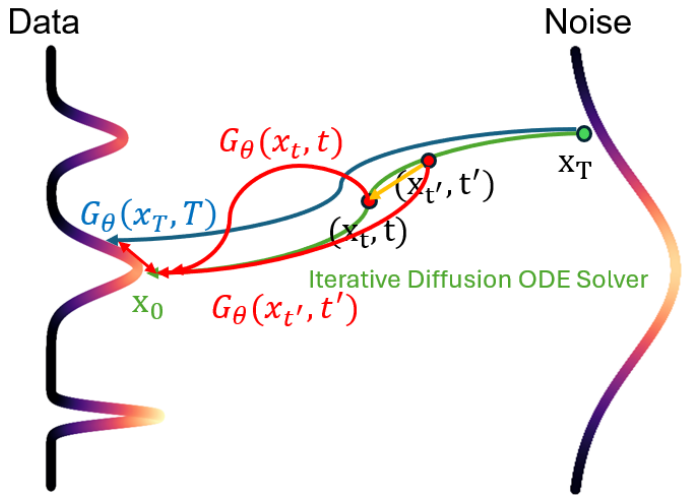
Unable training, scalability issue for large-scale text-to-image generation

Consistency Model: $\mathrm{L}(\theta) = \mathbb{E}_{\mathrm{x_t}}[\|\mathrm{G_\theta}(\mathrm{x_t}, \mathrm{t}) - \mathrm{G}_\theta(\mathrm{x_{t'}}, \mathrm{t'})\|]$



Knowledge Distillation: $\mathrm{L}(\theta) = \mathbb{E}_{\mathrm{x_T}}[\|\mathrm{G_\theta}(\mathrm{x_T}, \mathrm{T}) - \mathrm{x_0}\|]$

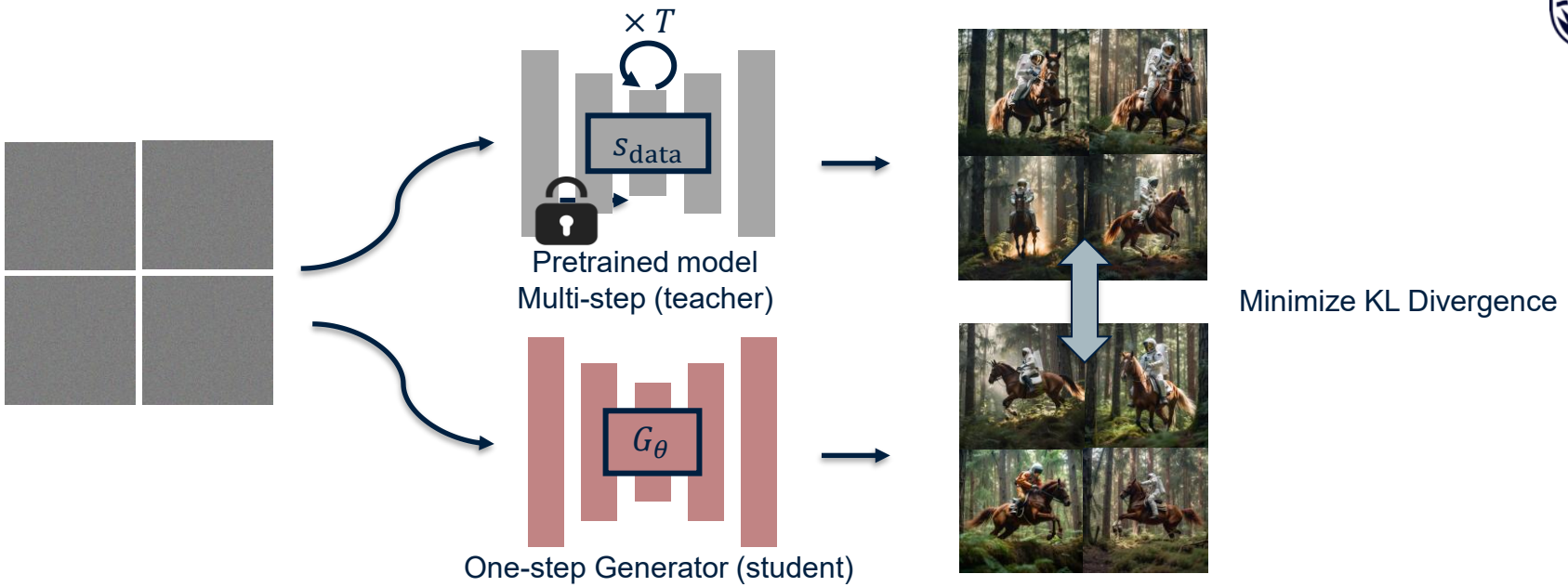




Song et al. Consistency Models. ICML 2023. Luhman et al. Knowledge Distillation. ArXiv 2021

Distribution Matching

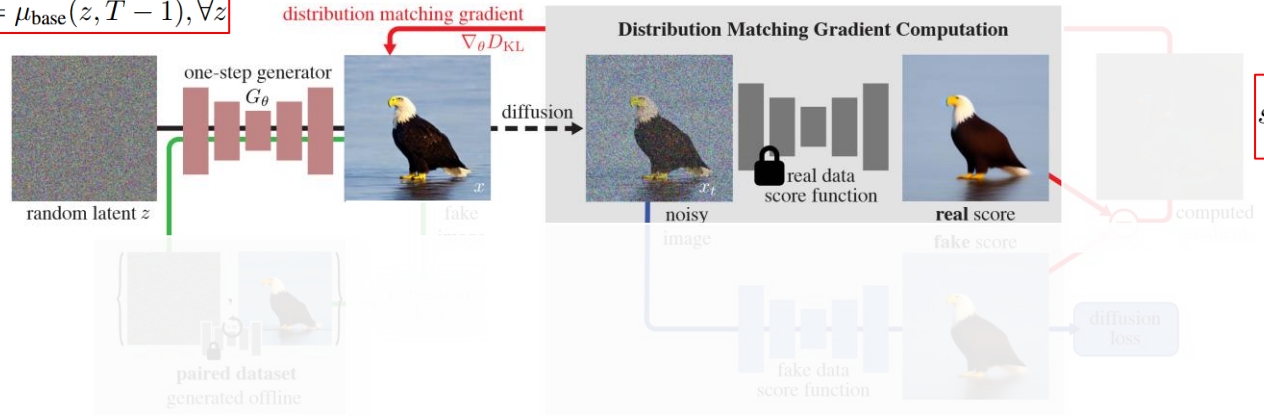
Match teacher model at distribution level instead



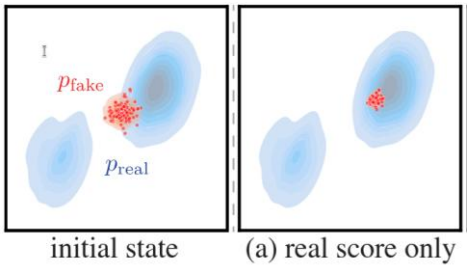
One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

Distribution Matching Distillation (DMD) – Real Score

$$G_\theta(z) = \mu_{\text{base}}(z, T-1), \forall z$$



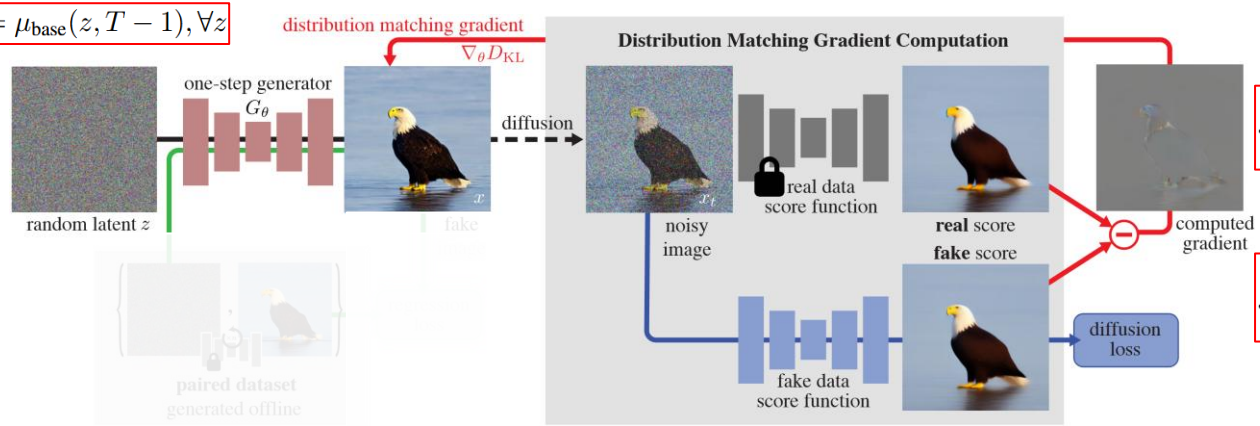
$$s_{\text{real}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{base}}(x_t, t)}{\sigma_t^2}$$



One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

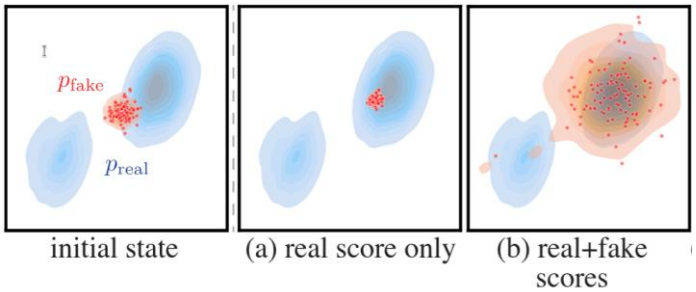
Distribution Matching Distillation (DMD) – Fake Score

$$G_\theta(z) = \mu_{\text{base}}(z, T-1), \forall z$$



$$s_{\text{real}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{base}}(x_t, t)}{\sigma_t^2}$$

$$s_{\text{fake}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{fake}}^{\phi}(x_t, t)}{\sigma_t^2}$$

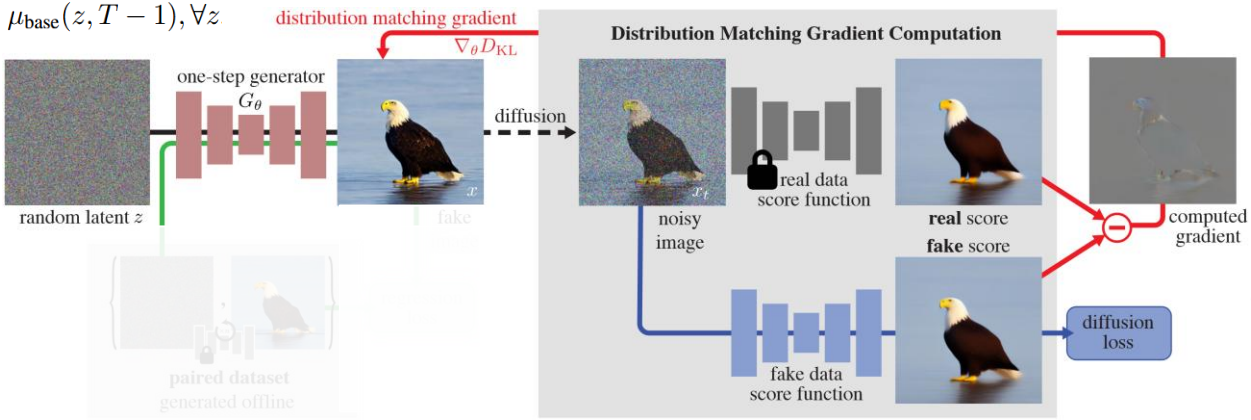


One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

Distribution Matching Distillation (DMD) – KL gradient update

$$D_{KL}\left(p_{\text{fake}} \,\|\, p_{\text{real}}\right) = \mathop{\mathbb{E}}_{x\sim p_{\text{fake}}}\left(\log\left(\frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)}\right)\right)$$
$$= \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I})\\ x=G_\theta(z)}} -\big(\log\, p_{\text{real}}(x) - \log\, p_{\text{fake}}(x)\big)$$

$$G_\theta(z) = \mu_{\text{base}}(z, T-1), \forall z$$



$$s_{\text{real}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{base}}(x_t, t)}{\sigma_t^2}$$

$$s_{\text{fake}}(x_t, t) = -\frac{x_t - \alpha_t \mu^{\phi}_{\text{fake}}(x_t, t)}{\sigma_t^2}$$

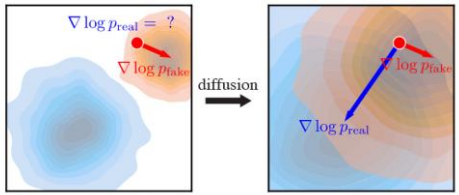
Gradient update using approximate scores

$$\nabla_\theta D_{KL} = \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I})\\ x=G_\theta(z)}}\left[-\big(s_{\text{real}}(x) - s_{\text{fake}}(x)\big)\frac{dG}{d\theta}\right] \quad s_{\text{real}}(x) = \nabla_x \log p_{\text{real}}(x),\ s_{\text{fake}}(x) = \nabla_x \log p_{\text{fake}}(x)$$

Distribution matching gradient update, with perturbed samples x_t, weighting factor ω_t

$$\nabla_\theta D_{KL} \simeq \mathop{\mathbb{E}}_{z,t,x,x_t}\left[w_t \alpha_t \big(s_{\text{fake}}(x_t, t) - s_{\text{real}}(x_t, t)\big)\frac{dG}{d\theta}\right] \quad w_t = \frac{\sigma_t^2}{\alpha_t}\frac{CS}{||\mu_{\text{base}}(x_t, t) - x||_1}$$

S: num of spatial location
C: num of channels



(a) for unperturbed distributions, both scores may not be defined simulateneously everywhere

(b) after diffusion, the distributions overlap, making our objective well-defined

One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

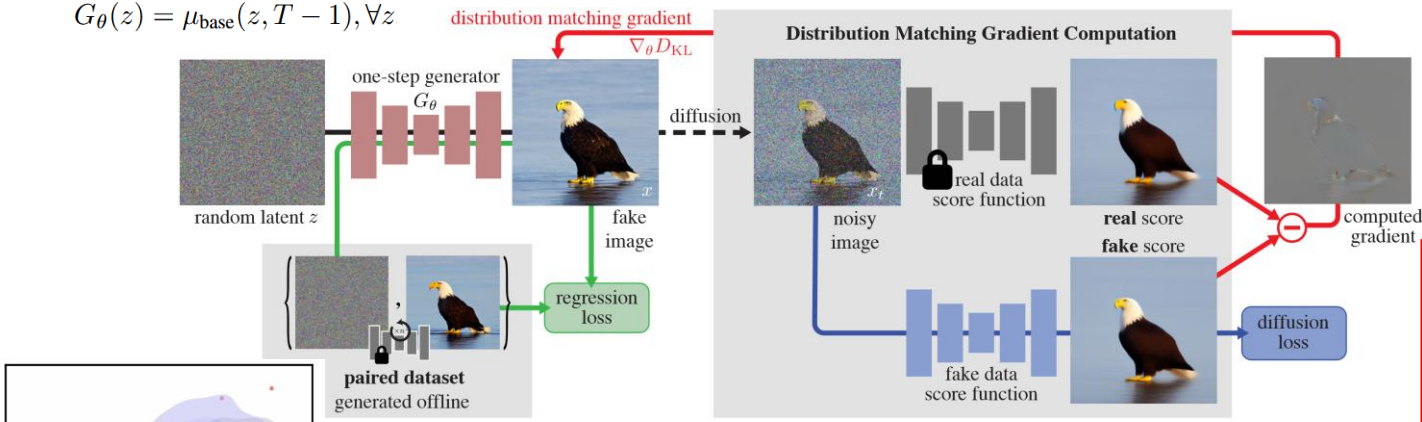
Distribution Matching Distillation (DMD) – Learning Objectives

$$D_{KL}\left(p_{\text{fake}} \parallel p_{\text{real}}\right) = \mathop{\mathbb{E}}_{x\sim p_{\text{fake}}}\left(\log\left(\frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)}\right)\right)$$
$$= \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I})\\ x=G_\theta(z)}} -\left(\log\, p_{\text{real}}(x) - \log\, p_{\text{fake}}(x)\right)$$

Gradient update using approximate scores

$$\nabla_\theta D_{KL} = \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I})\\ x=G_\theta(z)}}\left[-\left(s_{\text{real}}(x) - s_{\text{fake}}(x)\right)\frac{dG}{d\theta}\right] \quad s_{\text{real}}(x) = \nabla_x \log p_{\text{real}}(x),\ s_{\text{fake}}(x) = \nabla_x \log p_{\text{fake}}(x)$$

$$G_\theta(z) = \mu_{\text{base}}(z, T-1), \forall z$$

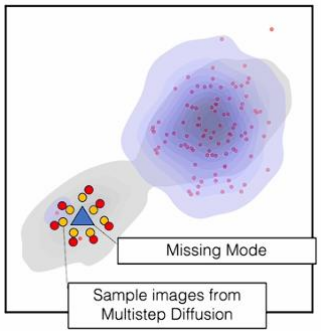


$$s_{\text{real}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{base}}(x_t, t)}{\sigma_t^2}$$

$$s_{\text{fake}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{fake}}^{\phi}(x_t, t)}{\sigma_t^2}$$

Dynamically-learned fake score
Standard denoising objective

$$\mathcal{L}_{\text{denoise}}^{\phi} = ||\mu_{\text{fake}}^{\phi}(x_t, t) - x_0||_2^2$$



Paired dataset $(z, y)\sim\mathcal{D}, y = \mu_{\text{base}}(z)$
Final Objective:

$$D_{KL} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \qquad \mathcal{L}_{\text{reg}} = \mathop{\mathbb{E}}_{(z,y)\sim\mathcal{D}} \ell(G_\theta(z), y)$$

Distance function ℓ: Learned Perceptual Image Patch Similarity (LPIPS)

One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

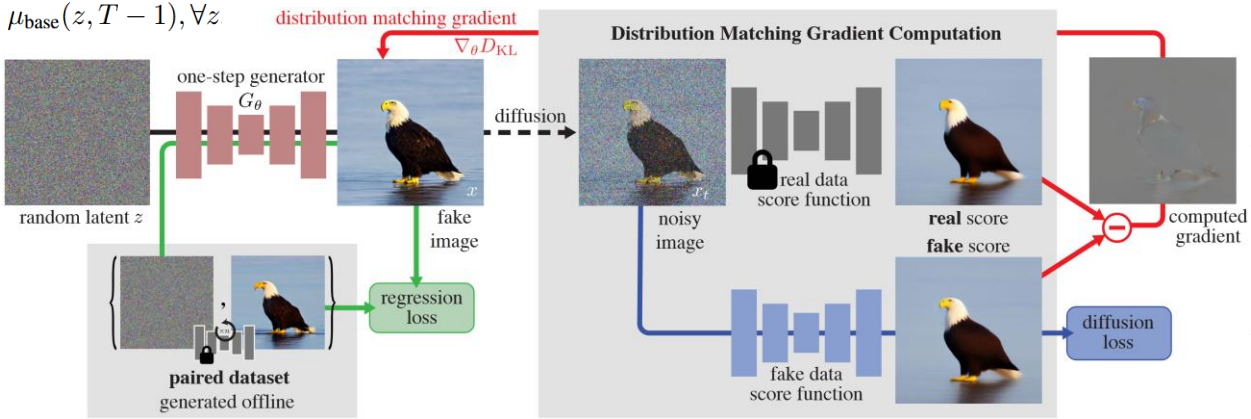
Distribution Matching Distillation (DMD)

$$D_{KL}\left(p_{\text{fake}} \parallel p_{\text{real}}\right) = \mathop{\mathbb{E}}_{x\sim p_{\text{fake}}}\left(\log\left(\frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)}\right)\right)$$
$$= \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I})\\ x=G_\theta(z)}} -\left(\log\, p_{\text{real}}(x) - \log\, p_{\text{fake}}(x)\right)$$

Gradient update using approximate scores

$$\nabla_\theta D_{KL} = \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I})\\ x=G_\theta(z)}}\left[-\left(s_{\text{real}}(x) - s_{\text{fake}}(x)\right)\frac{dG}{d\theta}\right] \quad s_{\text{real}}(x) = \nabla_x \log p_{\text{real}}(x),\ s_{\text{fake}}(x) = \nabla_x \log p_{\text{fake}}(x)$$

$$G_\theta(z) = \mu_{\text{base}}(z, T-1), \forall z$$



$$s_{\text{real}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{base}}(x_t, t)}{\sigma_t^2}$$

$$s_{\text{fake}}(x_t, t) = -\frac{x_t - \alpha_t \mu^{\phi}_{\text{fake}}(x_t, t)}{\sigma_t^2}$$

Distribution matching gradient update, with perturbed samples x_t, weighting factor ω_t

$$\nabla_\theta D_{KL} \simeq \mathop{\mathbb{E}}_{z,t,x,x_t}\left[w_t \alpha_t \left(s_{\text{fake}}(x_t, t) - s_{\text{real}}(x_t, t)\right)\frac{dG}{d\theta}\right] \quad w_t = \frac{\sigma_t^2}{\alpha_t}\frac{CS}{||\mu_{\text{base}}(x_t,t) - x||_1}$$

S: num of spatial location
C: num of channels

Final Objective

$$D_{KL} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \qquad \mathcal{L}_{\text{reg}} = \mathop{\mathbb{E}}_{(z,y)\sim\mathcal{D}} \ell(G_\theta(z), y)$$

One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

Derivation for Distribution Matching Gradient

Diffused sample $x_t \sim q(x_t|x)$ is obtained by adding noise to one-step generator output $x = G_\theta(z)$ at diffusion time step t: $q_t(x_t|x) \sim \mathcal{N}(\alpha_t x; \sigma_t^2 \mathrm{I})$

$$D_{KL}\left(p_{\text{fake}} \parallel p_{\text{real}}\right) = \mathop{\mathbb{E}}_{x \sim p_{\text{fake}}} \left(\log \left(\frac{p_{\text{fake}}(x)}{p_{\text{real}}(x)} \right) \right)$$

$$= \mathop{\mathbb{E}}_{\substack{z \sim \mathcal{N}(0;\mathbf{I}) \\ x = G_\theta(z)}} -\big(\log\ p_{\text{real}}(x) - \log\ p_{\text{fake}}(x) \big)$$

$$s_{\text{real}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{base}}(x_t, t)}{\sigma_t^2}$$

$$s_{\text{fake}}(x_t, t) = -\frac{x_t - \alpha_t \mu_{\text{fake}}^{\phi}(x_t, t)}{\sigma_t^2}$$

$$\nabla_\theta D_{KL} \simeq \mathop{\mathbb{E}}_{z,t,x,x_t} \left[w_t \big(s_{\text{fake}}(x_t, t) - s_{\text{real}}(x_t, t) \big) \frac{\partial x_t}{\partial \theta} \right]$$

$$= \mathop{\mathbb{E}}_{z,t,x,x_t} \left[w_t \big(s_{\text{fake}}(x_t, t) - s_{\text{real}}(x_t, t) \big) \frac{\partial x_t}{\partial G_\theta(z)} \frac{\partial G_\theta(z)}{\partial \theta} \right]$$

$$= \mathop{\mathbb{E}}_{z,t,x,x_t} \left[w_t \big(s_{\text{fake}}(x_t, t) - s_{\text{real}}(x_t, t) \big) \frac{\partial x_t}{\partial x} \frac{\partial G_\theta(z)}{\partial \theta} \right]$$

$$= \mathop{\mathbb{E}}_{z,t,x,x_t} \left[w_t \alpha_t \big(s_{\text{fake}}(x_t, t) - s_{\text{real}}(x_t, t) \big) \frac{dG}{d\theta} \right] \tag{10}$$

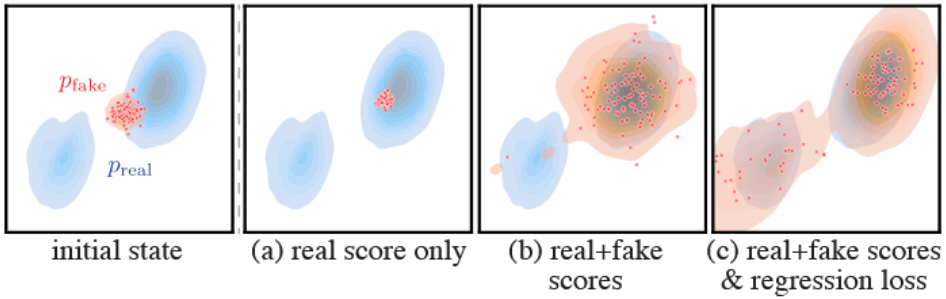
One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

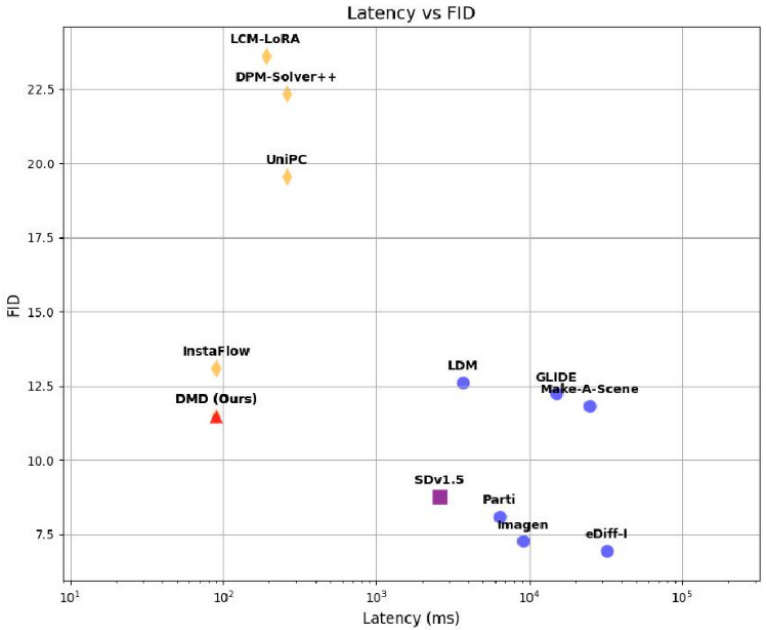
Distillating at Distribution Level

Distribution matching is easier than pair-wise mapping

Benchmarks: CIFAR-10, ImageNet

Text-to-Image Dataset: LAION-Aesthetic-6+

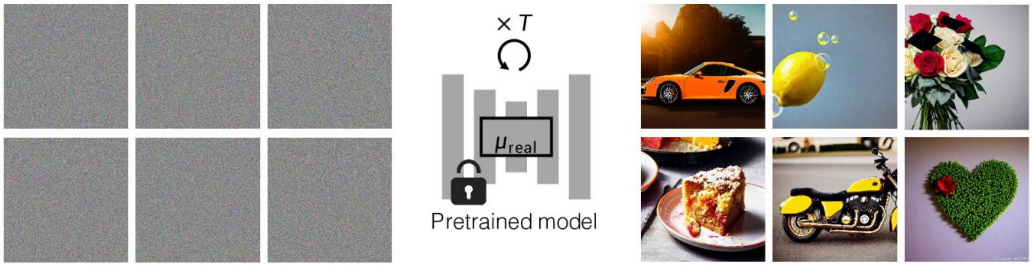




One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

Issue with DMD

- Costly for text-to-image synthesis
  - SDXL noise-image pair takes 5 secs; 700 A100 days to cover 12 million prompts in the LAION 6.0 dataset
  - Not generalizable to changing conditions (e.g. edge maps, depth maps, style, or different guidance scale)
- Performance bounder by the teacher diffusion model and specific ODE samplers
  - Large regression loss results in dominant point-wise supervision, prone to artifacts from previous methods

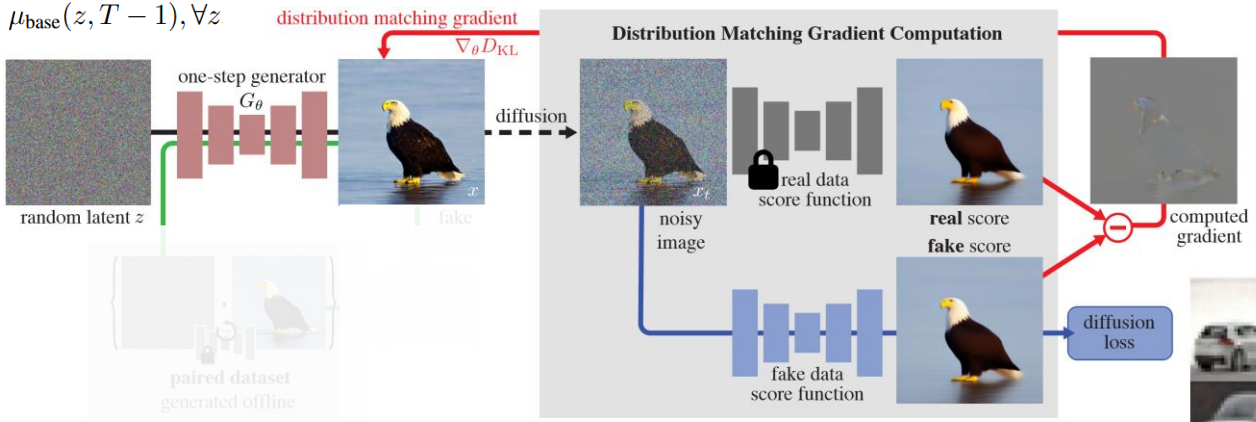


One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

DMD without regression loss

Removing regression yet ensuring training stability and prevent mode collapse!

$$G_\theta(z) = \mu_{\text{base}}(z, T-1), \forall z$$



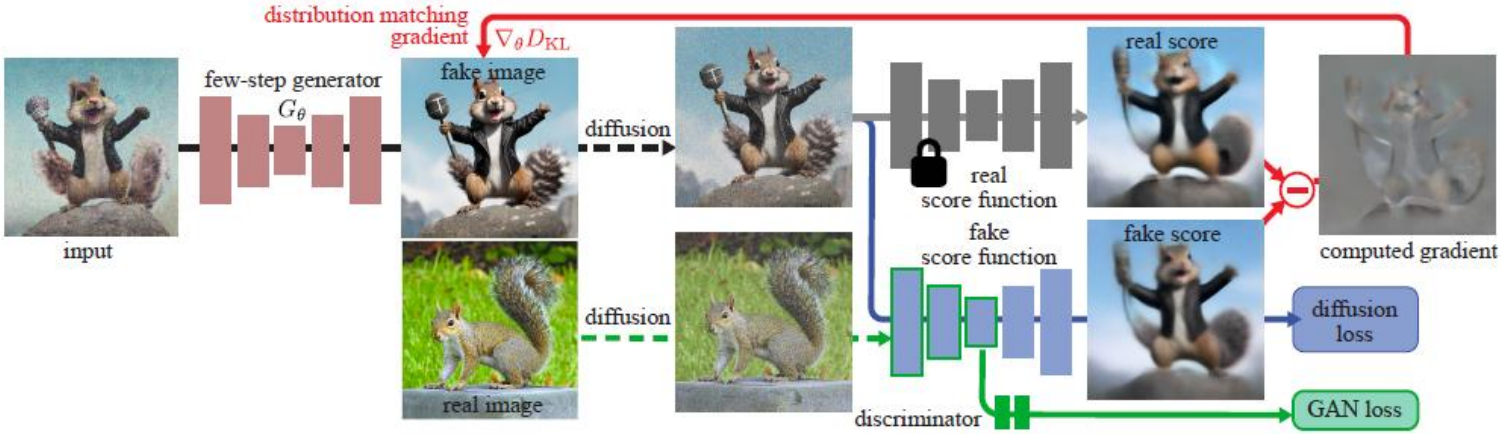
$$\nabla_\theta D_{KL} = \mathop{\mathbb{E}}_{\substack{z\sim\mathcal{N}(0;\mathbf{I}) \\ x=G_\theta(z)}} \Big[-\big(s_{\text{real}}(x) - s_{\text{fake}}(x)\big) \frac{dG}{d\theta} \Big]$$

$$D_{KL} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}$$

without regression loss

One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

Improved Distribution Matching Distillation (DMD2)

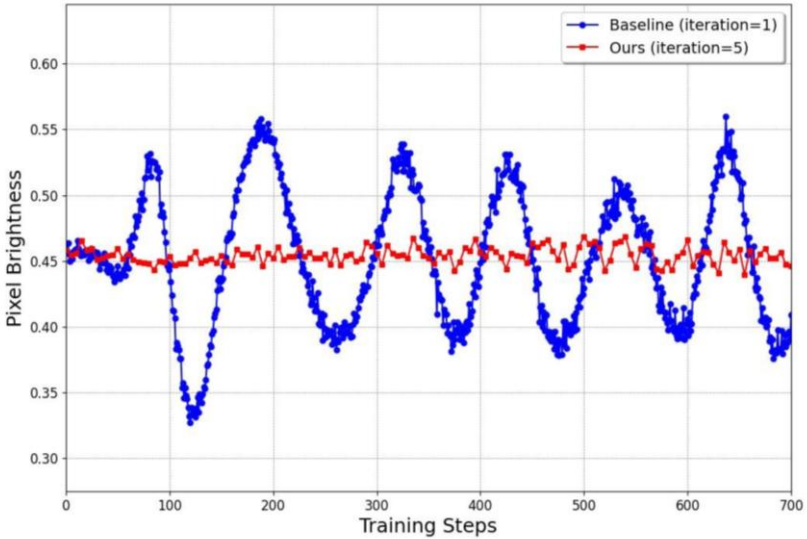


Model Pipeline for DMD2

Improved Distribution Matching Distillation for Fast Image Synthesis. Tianwei Yin, et al. NeruIPS (2024).

Training stabilization following regression removal

- Removing the regression loss introduces training instability
- Solution: Two Time-scale Update Rule for generator and fake score network
- One generator update for every 5 fake score model updates
- Pixel brightness variation depicted in the figure



Surpassing the teacher using GAN loss

- So far, the model is on par with teacher, without costly dataset (details in experiments)
- Real data is not used in training
- Proposed solution: Adding a GAN loss

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x\sim p_{\text{real}},t\sim[0,T]}[\log D(F(x,t))] + \mathbb{E}_{z\sim p_{\text{noise}},t\sim[0,T]}[-\log(D(F(G_\theta(z),t)))],$$

- GAN classifier: On top of middle layer of fake score diffusion model

Multi-step (few-step) generator

- Larger scale models like SDXL remain challenging to distill
- Solution: extend DMD to support multi-step sampling.
- A fixed schedule for both training and inference.
- During inference, starting with noise $z_0 \sim \mathcal{N}(0, \mathbf{I})$, alternate between denoising updates $\hat{x}_{t_i} = G_\theta(x_{t_i}, t_i)$, and forward diffusion $x_{t_{i+1}} = \alpha_{t_{i+1}} \hat{x}_{t_i} + \sigma_{t_{i+1}} \epsilon$
- Model uses the schedule: 999, 749, 499, 249

Left: Forward diffusion of real images during training
Right: Simulating the backwards process during training, to reach train-test alignment.

Experiments

Class-Conditional Image Generation

Comparison of one-step generator with baselines

On ImageNet-64 × 64

- Fréchet Inception Distance (FID) metric used

To measure quality

- Success attributed to removing regression loss

Method	# Fwd Pass (↓)	FID (↓)
BigGAN-deep [65]	1	4.06
ADM [66]	250	2.07
RIN [67]	1000	1.23
StyleGAN-XL [35]	1	1.52
Progress. Distill. [10]	1	15.39
DFNO [68]	1	7.83
BOOT [20]	1	16.30
TRACT [33]	1	7.43
Meng et al. [13]	1	7.54
Diff-Instruct [44]	1	5.57
Consistency Model [9]	1	6.20
iCT-deep [12]	1	3.25
CTM [26]	1	1.92
DMD [22]	1	2.62
DMD2 (Ours)	1	**1.51**
+longer training (Ours)	1	**1.28**
EDM (Teacher, ODE) [52]	511	2.32
EDM (Teacher, SDE) [52]	511	1.36

Image quality comparison on ImageNet-64 × 64

Text-to-Image Synthesis

- Competitve CLIP score, and superior FID and Patch FID scores

Method	# Fwd Pass (↓)	FID (↓)	Patch FID (↓)	CLIP (↑)
LCM-SDXL [32]	1	81.62	154.40	0.275
	4	22.16	33.92	0.317
SDXL-Turbo [23]	1	24.57	23.94	**0.337**
	4	23.19	23.27	0.334
SDXL Lightning [27]	1	23.92	31.65	0.316
	4	24.46	24.56	0.323
DMD2 (Ours)	1	**19.01**	26.98	0.336
	4	19.32	**20.86**	0.332
SDXL Teacher, cfg=6 [57]	100	19.36	21.38	0.332
SDXL Teacher, cfg=8 [57]	100	20.39	23.21	0.335

Image quality comparison with SDXL on 10K prompts from COCO 2014.

Text-to-Image Synthesis

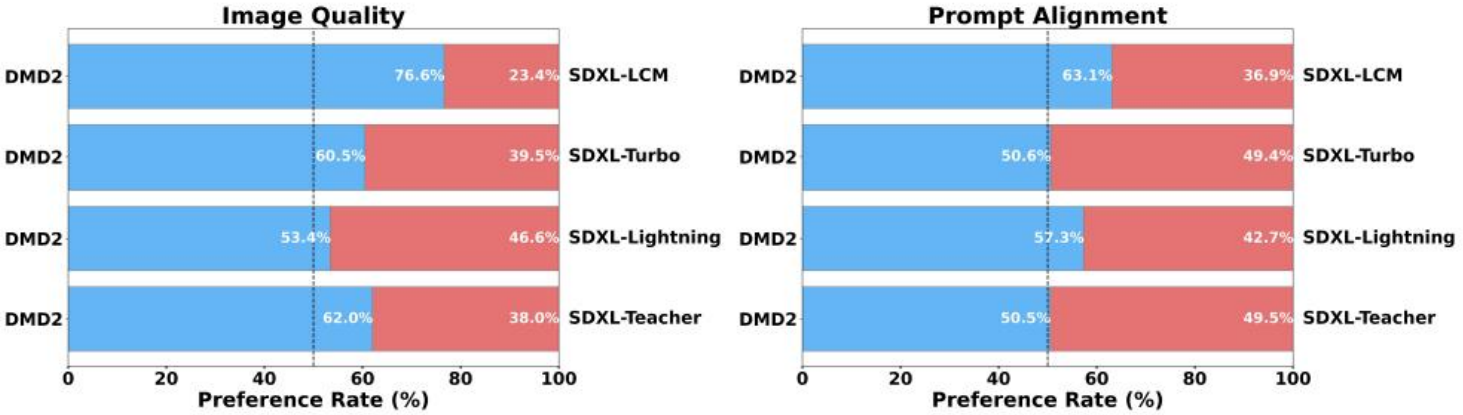
- Similar experiments for SDv1.5

Family	Method	Resolution (↑)	Latency (↓)	FID (↓)
Original, unaccelerated	DALL·E [77]	256	-	27.5
	DALL·E 2 [3]	256	-	10.39
	Parti-750M [69]	256	-	10.71
	Parti-3B [69]	256	6.4s	8.10
	Make-A-Scene [78]	256	25.0s	11.84
	GLIDE [79]	256	15.0s	12.24
	LDM [1]	256	3.7s	12.63
	Imagen [4]	256	9.1s	7.27
	eDiff-I [5]	256	32.0s	**6.95**
GANs	LAFITE [80]	256	0.02s	26.94
	StyleGAN-T [81]	512	0.10s	13.90
	GigaGAN [71]	512	0.13s	**9.09**
Accelerated diffusion	DPM++ (4 step) [50]	512	0.26s	22.36
	UniPC (4 step) [82]	512	0.26s	19.57
	LCM-LoRA (4 step) [32]	512	0.19s	23.62
	InstaFlow-0.9B [11]	512	0.09s	13.10
	SwiftBrush [45]	512	0.09s	16.67
	HiPA [83]	512	0.09s	13.91
	UFOGen [25]	512	0.09s	12.78
	SLAM (4 step) [18]	512	0.19s	10.06
	DMD [22]	512	0.09s	11.49
	DMD2 (Ours)	512	0.09s	**8.35**
Teacher	SDv1.5 (50 step, cfg=3, ODE) [1, 49]	512	2.59s	8.59
	SDv1.5 (200 step, cfg=2, SDE) [1, 46]	512	10.25s	7.21

Comparison with SDv1.5 on 30K prompts from COCO 2014.

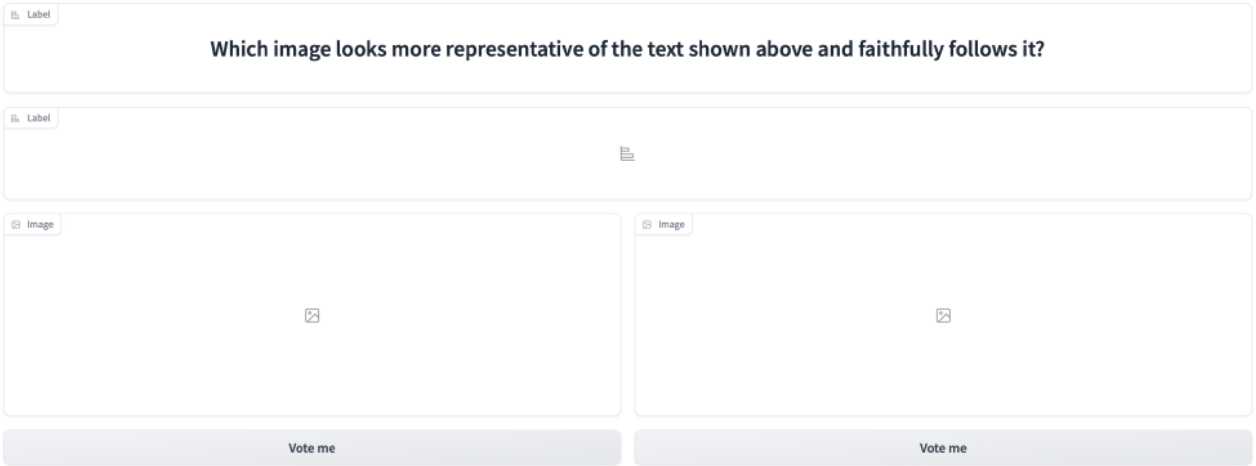
User studies

User study comparing distilled model with its teacher and competing distillation baselines. All distilled models use 4 sampling steps, the teacher uses 50.



User Study Details

- Images are presented in pairs to five random human evaluators
- 128 prompts from the LADD subset of PartiPrompts to create images



Text-to-Image Synthesis: Further Analysis

- Added LPIPS-based diversity score
- Four images are generated per prompt to calculate average pairwise LPIPS distance
- The LADD subset of PartiPrompts was used
- Slight diversity degradation

Method	# Fwd Pass (↓)	FID (↓)	Patch FID (↓)	CLIP (↑)	Diversity Score (↑)
LCM-SDXL [32]	4	22.16	33.92	0.317	0.61
SDXL-Turbo [23]	4	23.19	23.27	**0.334**	0.58
SDXL-Lightning [27]	4	24.46	24.56	0.323	**0.63**
DMD2 (Ours)	4	**19.32**	**20.86**	0.332	0.61
SDXL-Teacher, cfg=6 [57]	100	19.36	21.38	0.332	0.64
SDXL-Teacher, cfg=8 [57]	100	20.39	23.21	0.335	0.64

Ablation studies

- Ablation showing the effects of each component in Table 3
- TTUR fully mitigates removing the regression loss and thus costly noise-image pairs
- GAN and DMD terms should both be present for best outcome

DMD	No Regress.	TTUR	GAN	FID (↓)
✓				2.62
✓	✓			3.48
✓	✓	✓		2.61
✓	✓	✓	✓	**1.51**
			✓	2.56
		✓	✓	2.52

Ablation studies on ImageNet.

Method	FID (↓)	Patch FID (↓)	CLIP (↑)
w/o GAN	26.90	27.66	0.328
w/o Distribution Matching	**13.77**	27.96	0.307
w/o Backward Simulation	20.66	24.21	0.332
DMD2 (Ours)	19.32	**20.86**	**0.332**

Ablation studies with SDXL backbone on 10K prompts from COCO 2014.

Visual Comparison

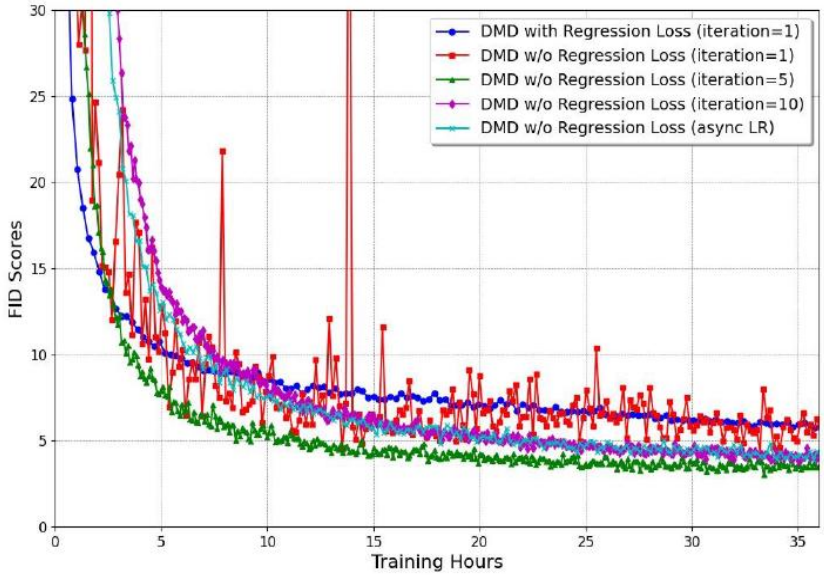
Visual comparison between DMD2, the SDXL teacher, and competing methods.
All models use identical noise and prompts. Distilled models use 4 sampling steps

Visual Ablation

- All images are generated using identical noise and text prompts.
- Omitting the GAN loss: oversaturated and overly smoothed images.

Two Time-scale Update Rule

Experiments to determine optimal time scaling factor

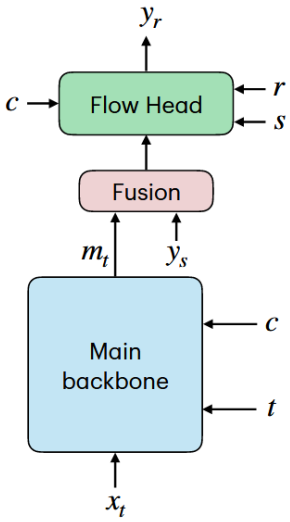


Visualization of FID score during training

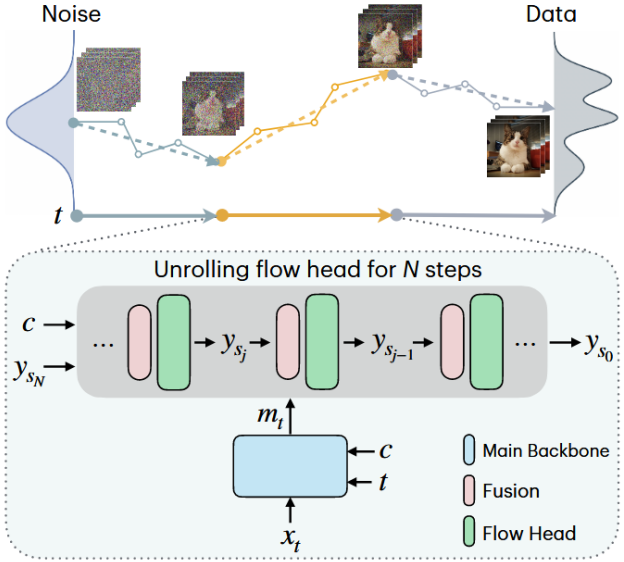
Discussion, Follow Up/Future Works

Transition Matching Distillation for Fast Video Generation

Distilling video diffusion models into efficient few-step generators. (1) Main backbone, comprising the majority of early layers, semantic representations at each outer transition step. (2) flow head, last few layers, leverages these representation to perform multiple inner flow updates.



(a) Decoupled architecture



(b) Transition process with flow head rollout

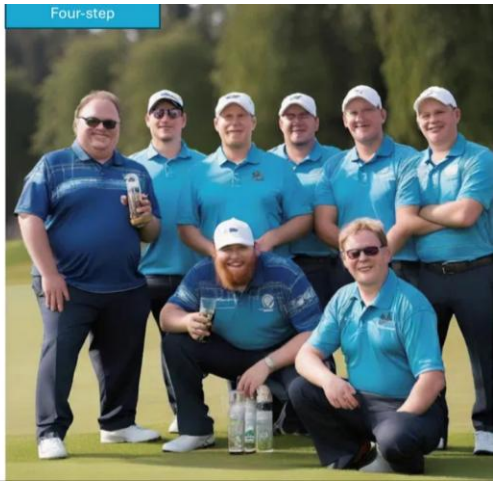
Algorithm 2 TMD student update step (simplified)

Given $\boldsymbol{x} \sim p_{\text{data}},\ \boldsymbol{x}_1 \sim \mathcal{N}(0, I),\ t_i \sim \text{Unif}(\{t_1, \ldots, t_M\})$
$\boldsymbol{x}_{t_i} = (1 - t_i)\boldsymbol{x} + t_i \boldsymbol{x}_1$ ▷ See Eq. (1)
$\boldsymbol{m} = \boldsymbol{m}_\theta(\boldsymbol{x}_{t_i}, t_i)$ ▷ Main backbone
if stage_one **then**
  $\boldsymbol{y} = \boldsymbol{x}_1 - \boldsymbol{x}$ ▷ See Eq. (3)
  $\boldsymbol{y}_1 \sim \mathcal{N}(0, I),\ (s, r) \sim p_{s,r}$
  $\boldsymbol{y}_s = (1 - s)\boldsymbol{y} + s\boldsymbol{y}_1$ ▷ See Eq. (5)
  $\boldsymbol{u} = \boldsymbol{u}_\theta(\boldsymbol{y}_s, s, r; \boldsymbol{m})$ ▷ Avg. velocity
  $\boldsymbol{v} = \boldsymbol{y}_1 - \boldsymbol{y}$ ▷ Conditional velocity
  $\mathcal{L} = \text{MeanFlow}(\boldsymbol{u}, \boldsymbol{v}, s, r)$ ▷ See Eq. (9)
else
  $\hat{\boldsymbol{x}} = \boldsymbol{x}_1 - \text{InnerFlow}(\boldsymbol{m})$ ▷ See Eq. (15) & Algorithm 1
  $\mathcal{L} = \text{VSD}(\hat{\boldsymbol{x}}) + \lambda \cdot \text{Discriminator}(\hat{\boldsymbol{x}})$ ▷ See Eq. (11)
$\theta = \text{step}(\theta, \nabla_\theta \mathcal{L})$ ▷ Gradient step

Transition Matching Distillation for Fast Video Generation. Weili Nie, et al. arXiv: 2601.09881v1 (2026). https://research.nvidia.com/labs/genair/tmd/

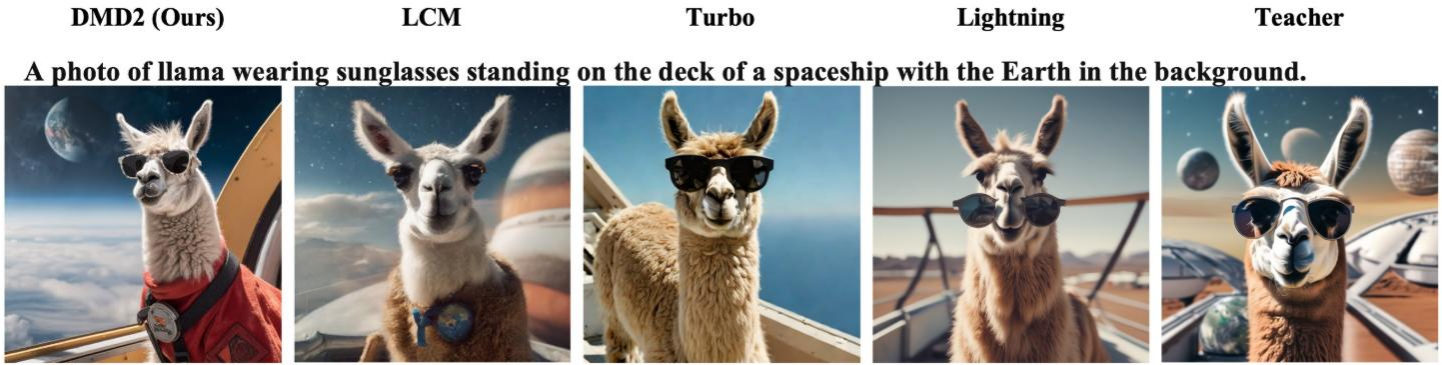
Limitation

- Degradation in image diversity compared to the teacher models
- Requires four steps to match SDXL model, no longer single step
- Fixed guidance scale during training limiting user flexibility
- Optimized for distribution matching while RLHF, RLVF could also improve performance
- Computationally expensive: 64 A100 GPUs 60 hours to distill SDXL



Conclusion

- Distribution Matching Distillation, diffusion distillation through matching the teacher and student denoising distribution
- Replicate teacher model performance with one- or few-step generators, reducing iteration by an order of magnitude
- Challenges remain in maintaining output diversity, and reductions have been observed
- DMD is a distillation approach. New tasks still requires diffusion fine-tuning followed by distillation



Reference

[1] Probabilistic Machine Learning. Kevin Murphy (2023).

[2] Song et al. Consistency Models. ICML (2023).

[3] Luhman et al. Knowledge Distillation. ArXiv (2021)

[4] One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. CVPR (2024).

[5] Improved Distribution Matching Distillation for Fast Image Synthesis. Tianwei Yin, et al. NeruIPS (2024).

[6] Transition Matching Distillation for Fast Video Generation. Weili Nie, et al. arXiv: 2601.09881v1 (2026).

THE UNIVERSITY OF BRITISH COLUMBIA