Refining Generative Process with **Discriminator Guidance** in Score-based Diffusion Models

Dongjun Kim*, Yeongmin Kim*,
Se Jung Kwon, Wanmo Kang, Il-Chul Moon
KAIST

* Equal contribution

Poor Sample, Why?

Before

- ✓ Discretization Error
- ✓ Score Estimation Error
- ✓ Prior Mismatch Error

Let's Reduce Score Estimation Error

Before

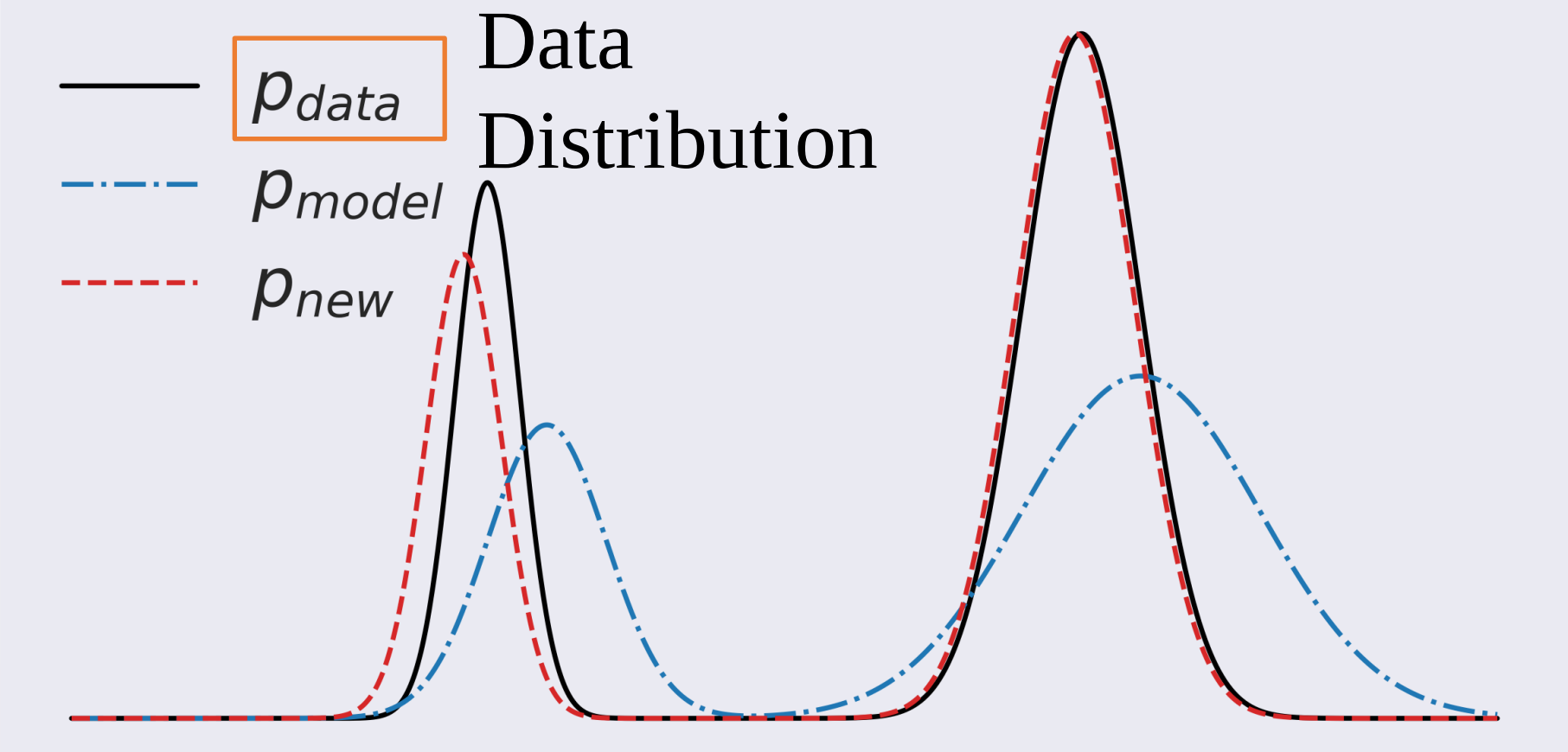
- ✓ Discretization Error
- ✓ **Score Estimation Error**
- ✓ Prior Mismatch Error

Our Method Improves Sample Quality

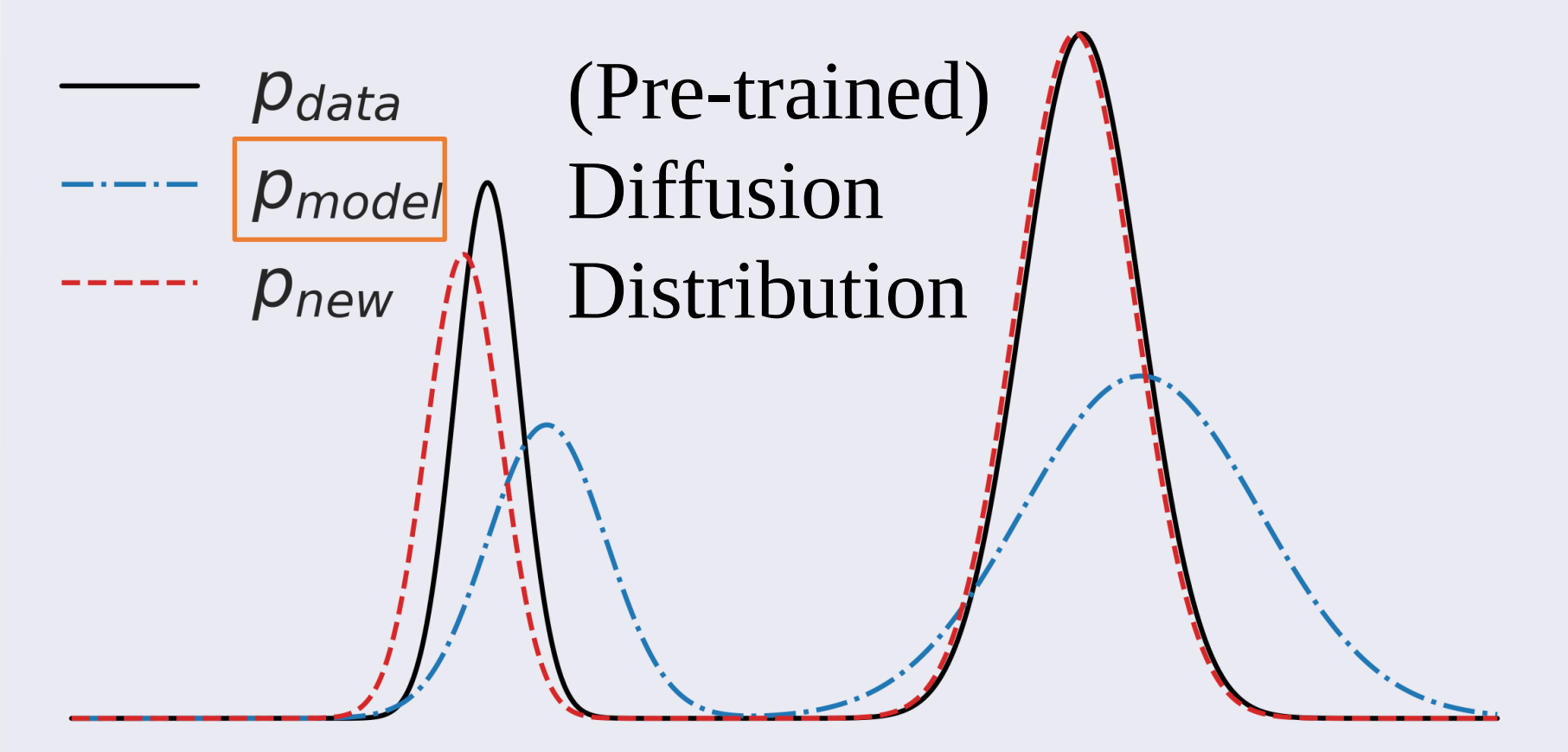
Before

After

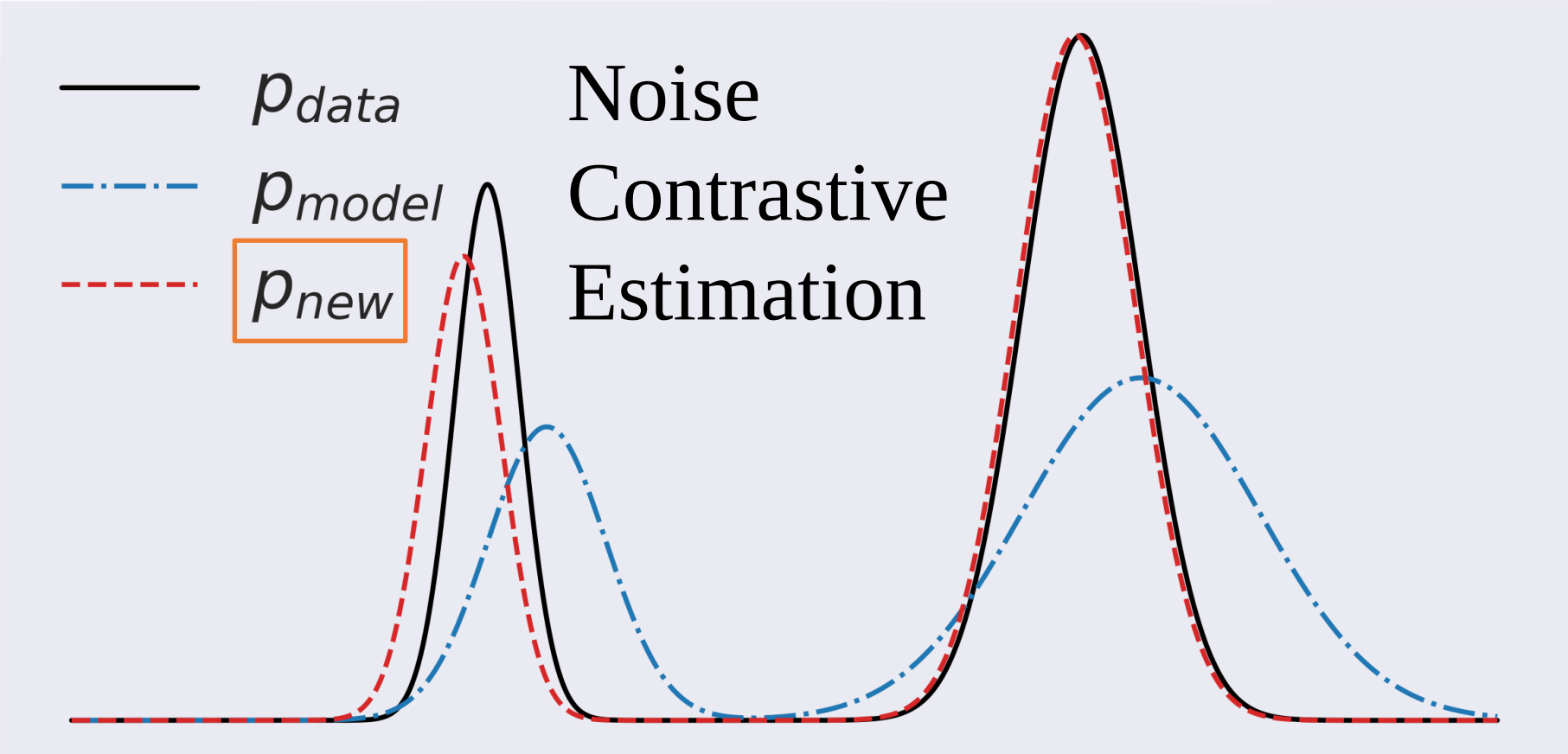
Noise Contrastive Estimation: Motivation



Noise Contrastive Estimation: Motivation

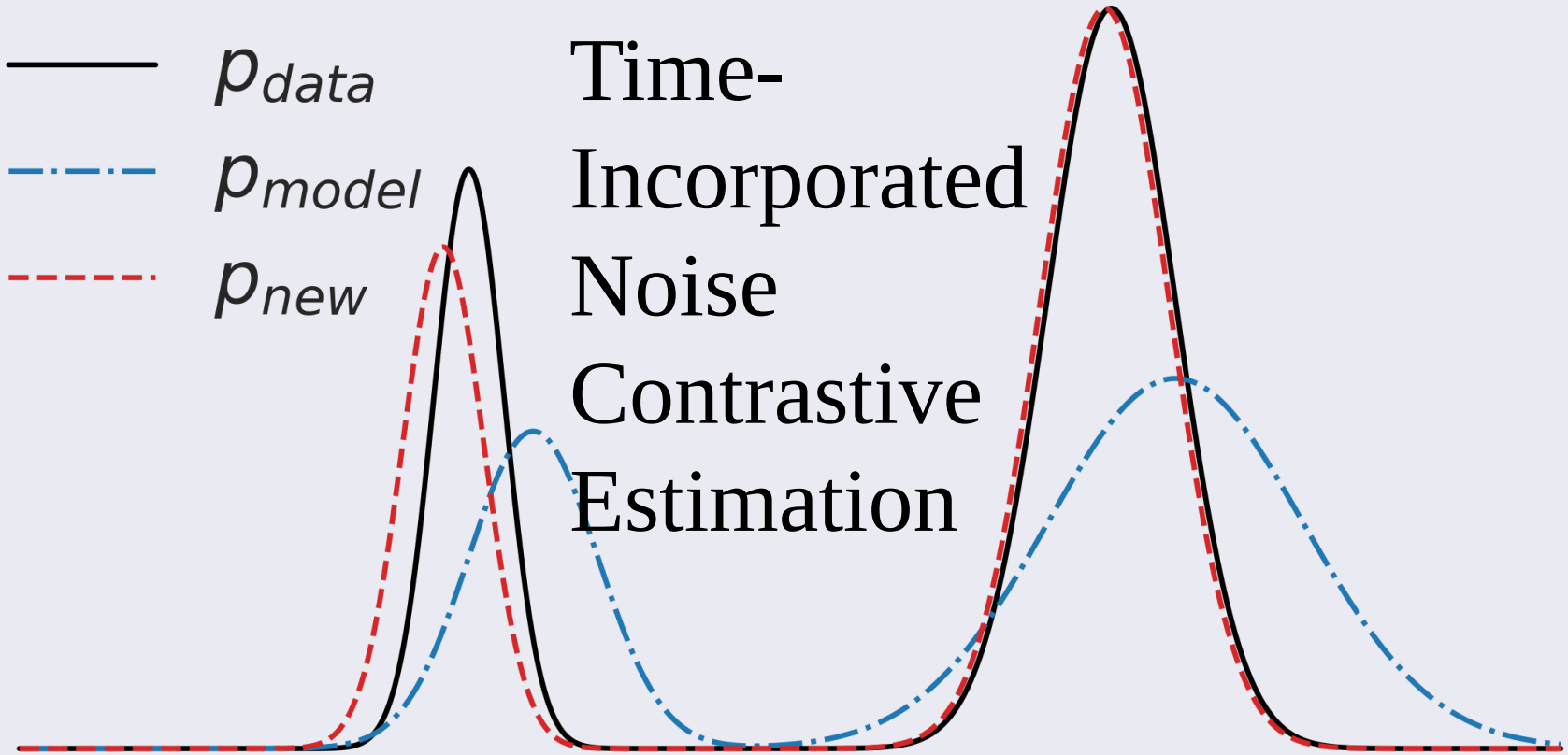


Noise Contrastive Estimation: Motivation



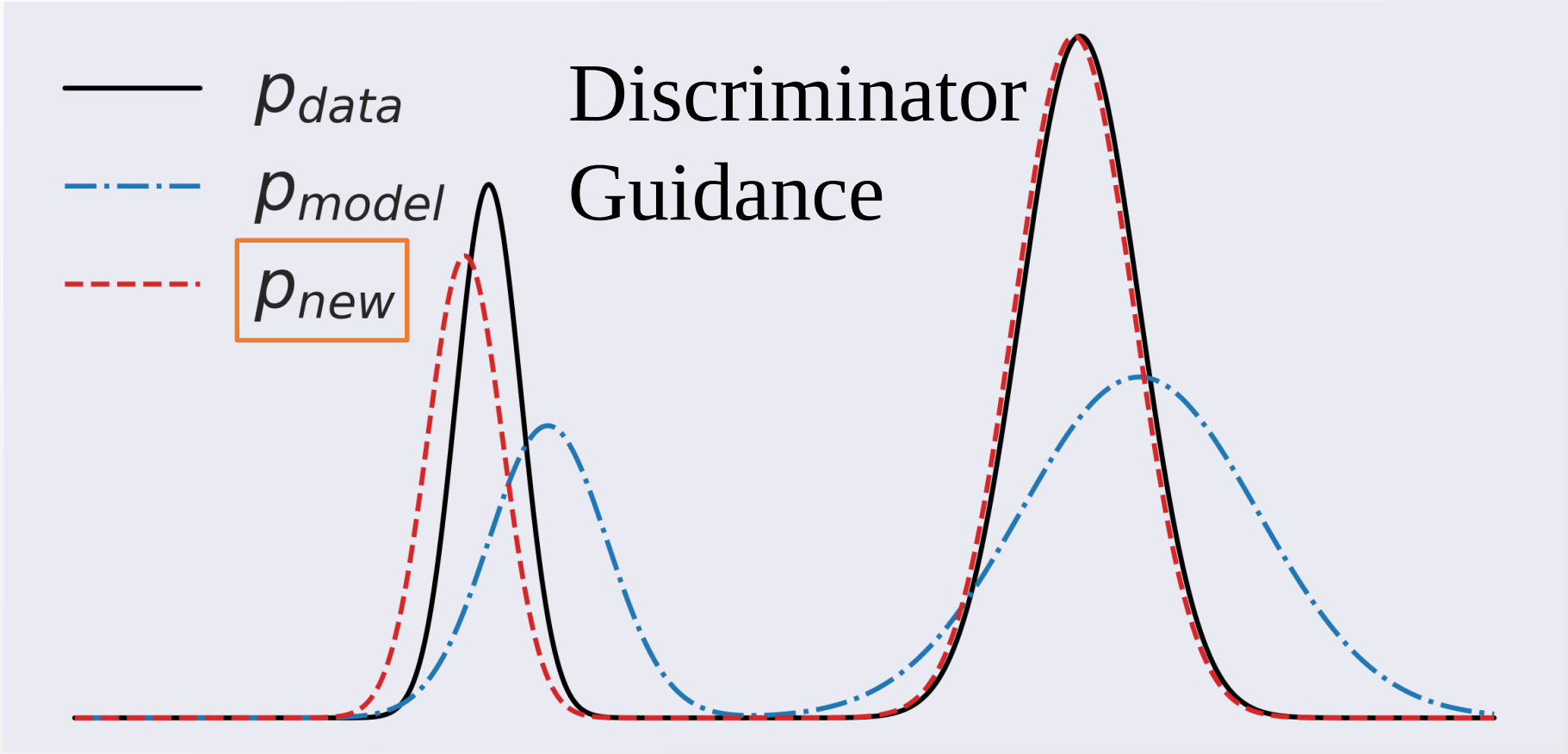
$$p_{new}(\mathbf{x}) \propto p_{model}(\mathbf{x}) \frac{d(\mathbf{x})}{1-d(\mathbf{x})}$$

Discriminator-Guided New Distribution



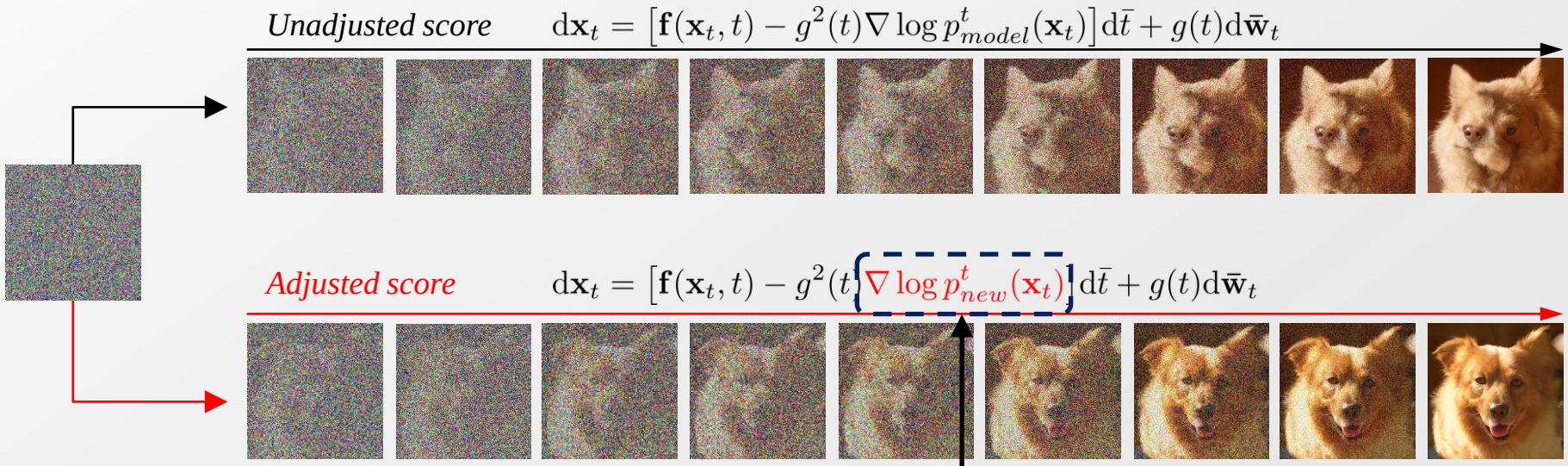
$$p^t_{new}(\mathbf{x}_t) \propto p^t_{model}(\mathbf{x}_t)\frac{d(\mathbf{x}_t,t)}{1-d(\mathbf{x}_t,t)}$$

Gradient + Logarithm on New Distribution



$$\nabla \log p^t_{new}(\mathbf{x}_t) = \underbrace{\nabla \log p^t_{model}(\mathbf{x}_t)}_{\text{Freeze}} + \underbrace{\nabla \log \frac{d(\mathbf{x}_t, t)}{1 - d(\mathbf{x}_t, t)}}_{\text{Disc. Guidance}}$$

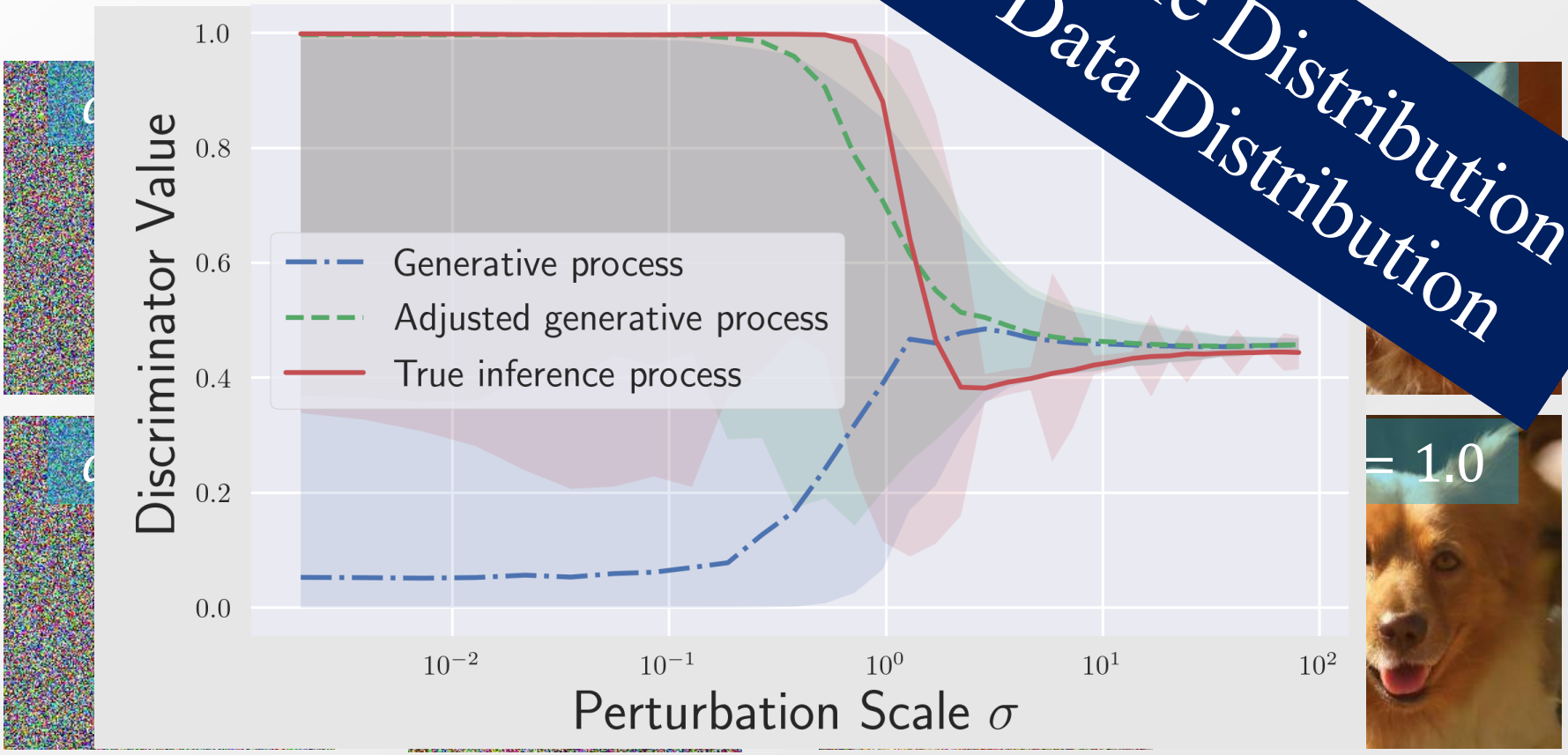
Sampling of Discriminator Guidance



$$\nabla \log p^t_{new}(\mathbf{x}_t) = \nabla \log p^t_{model}(\mathbf{x}_t) + \nabla \log \frac{d(\mathbf{x}_t, t)}{1 - d(\mathbf{x}_t, t)}$$

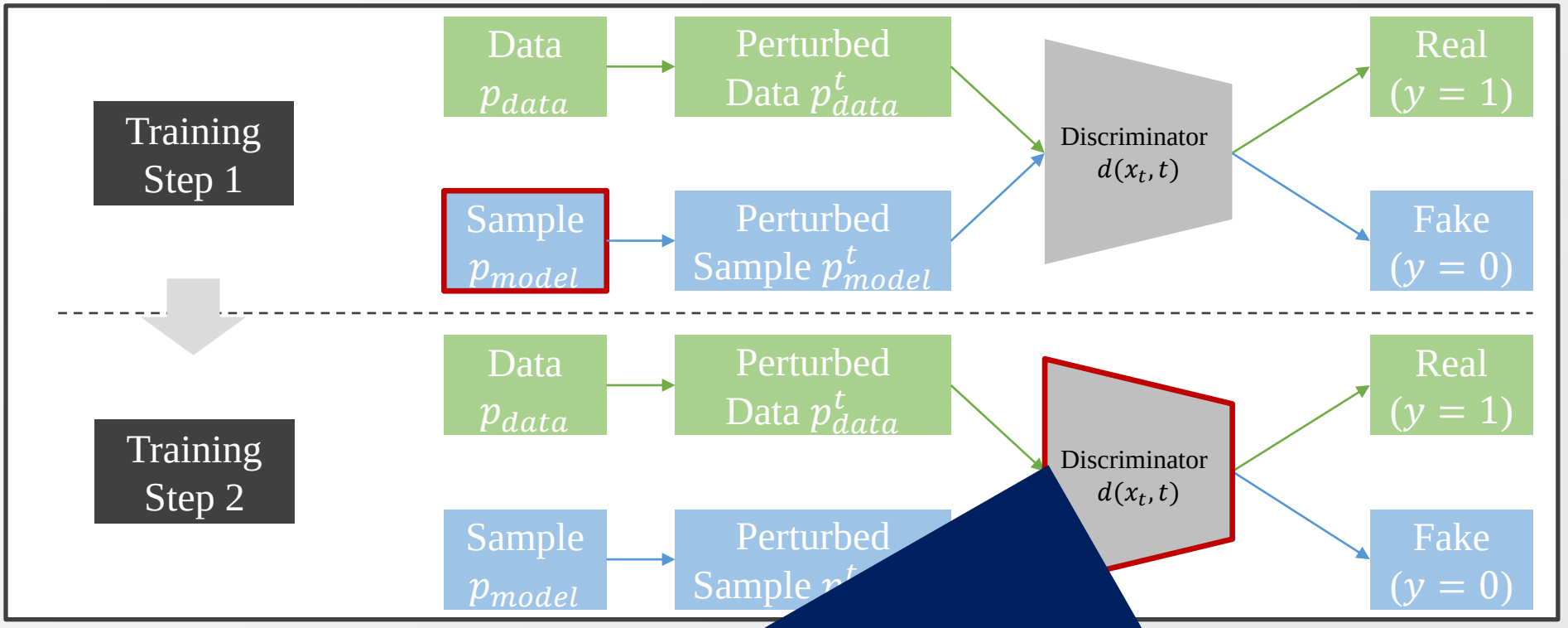
Meaning of Sampling

Sample Distribution = Data Distribution



$$p^t_{new}(\mathbf{x}_t) \uparrow\uparrow \propto p^t_{model}(\mathbf{x}_t) \uparrow\uparrow \frac{d(\mathbf{x}_t,t)}{1-d(\mathbf{x}_t,t)} \uparrow\uparrow$$

Learning Details



Cheap Training

Dataset (Model)	Score T…	Discriminator Training (Step 2)
CIFAR-10 (EDM)	…score eval (1hr)	1M discriminator eval (10min)
ImageNet 256x256 (ADI…	100M score eval	25.6M discriminator eval

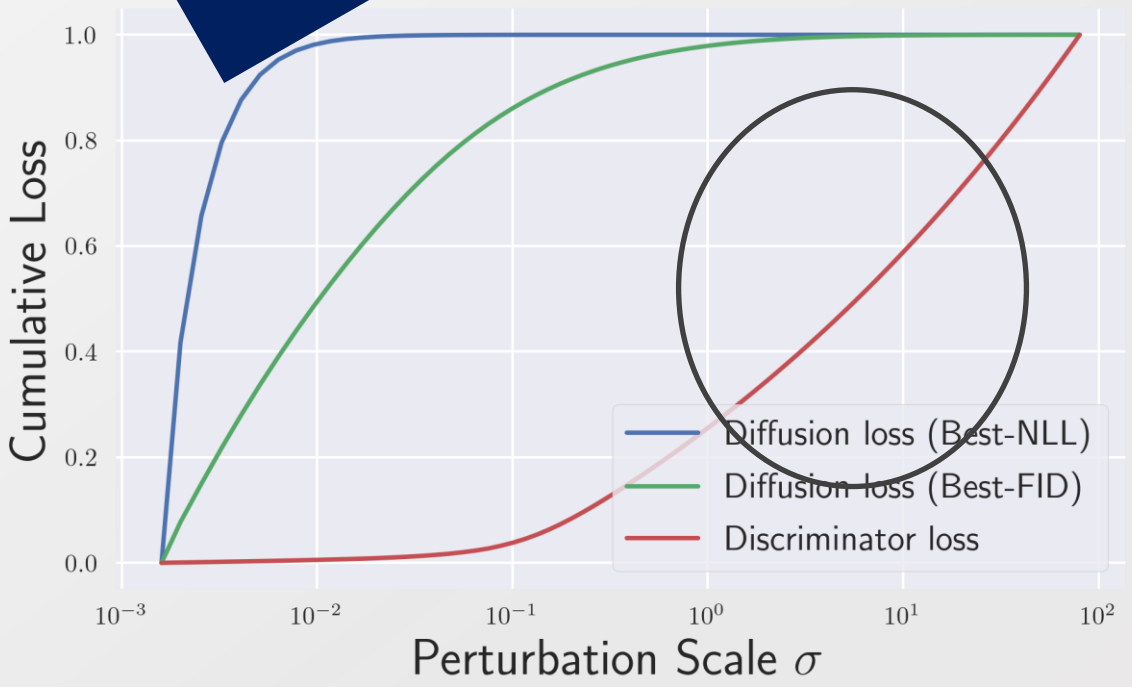
Training Characteristics

Binary

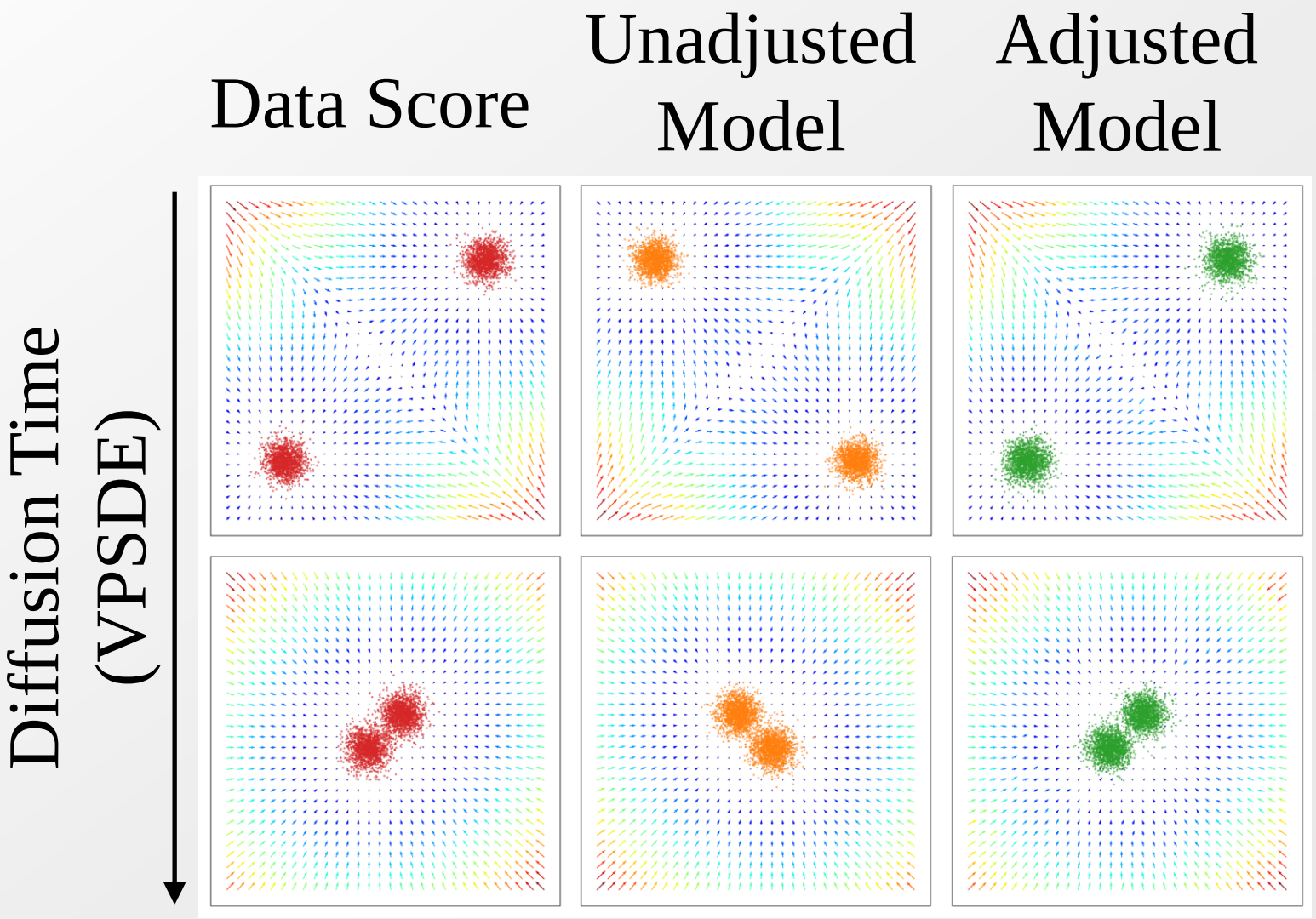
$$\mathcal{L}(\psi;\lambda) = \int_0^T \lambda$$

$$BCE_t(\psi) = \mathbb{E}_{p_{da}} \; \mathbb{E}_{p_{\theta_\infty}(\mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}\left[-\log\left(1 - d_\psi(\mathbf{x}_t, t)\right)\right]$$

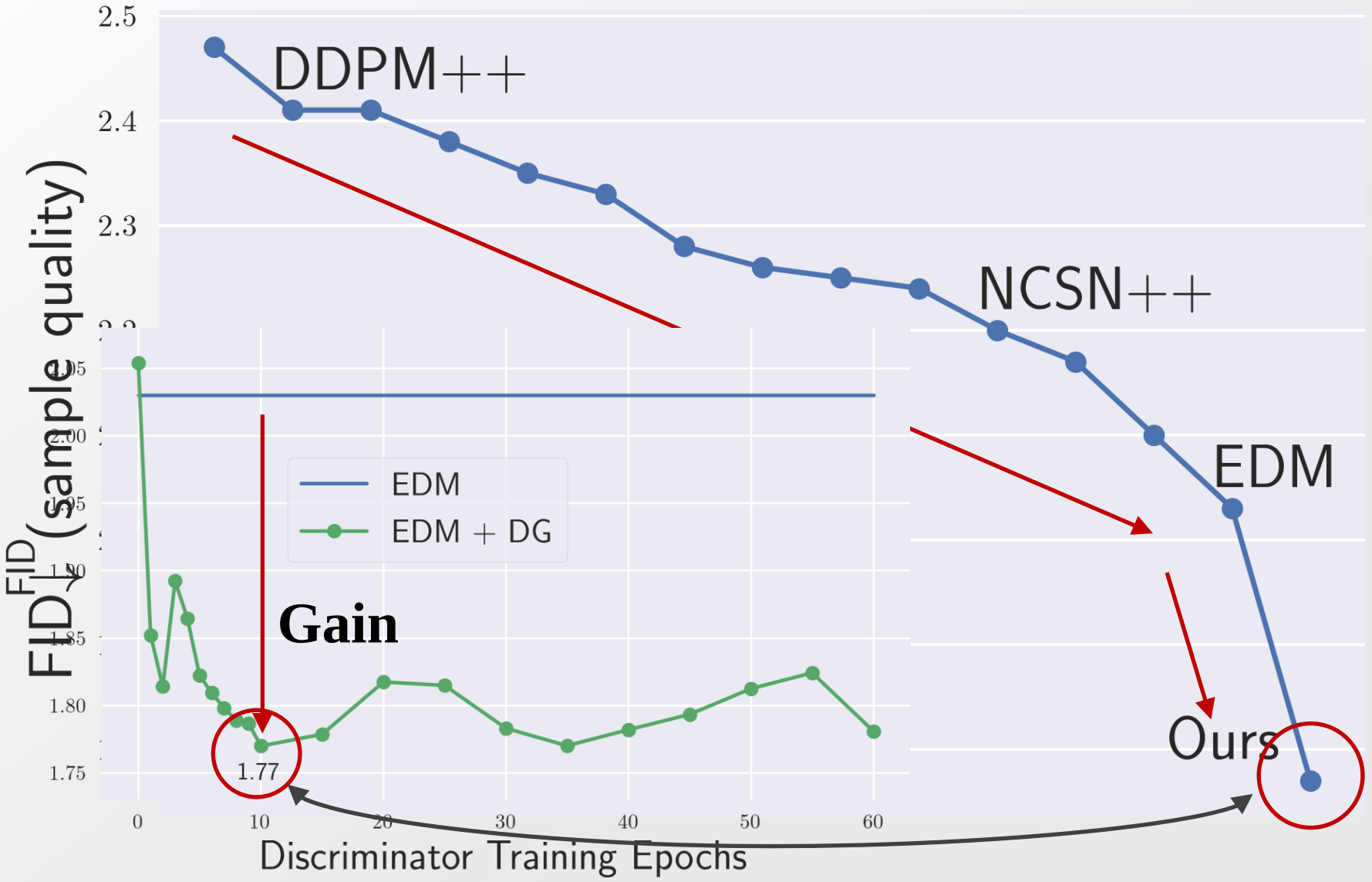
Adv. Sampling but Stable Training



Toy Experiment



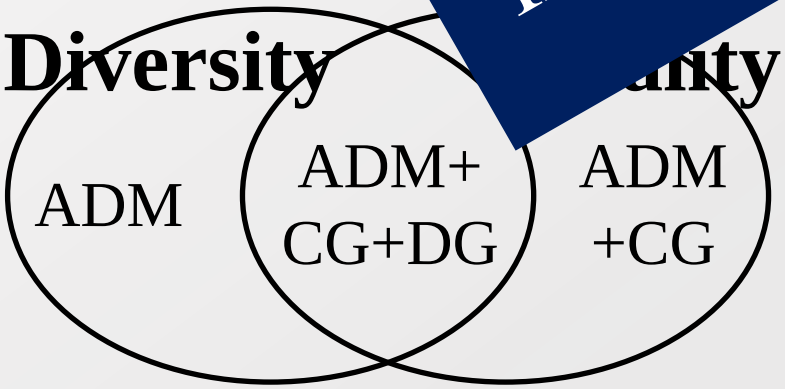
Record-breaking Quality in CIFAR-10



Record-breaking Quality&Diversity in ImageNet 256x256

Model	Diffusion Space	FID↓ (Quality)	Recall↑ (Diversity)
Validation Data		1.68	0.66
ADM	Data	10.94	0.63
ADM+CG	Data	4.59	0.52
ADM+CG+DG	Data	3.73	0.59
DiT-XL/2	Latent	9.62	
DiT-XL/2+CG	Latent	2.27	
DiT-XL/2+CG+DG	Latent		

Intra-Class Diversity



ADM

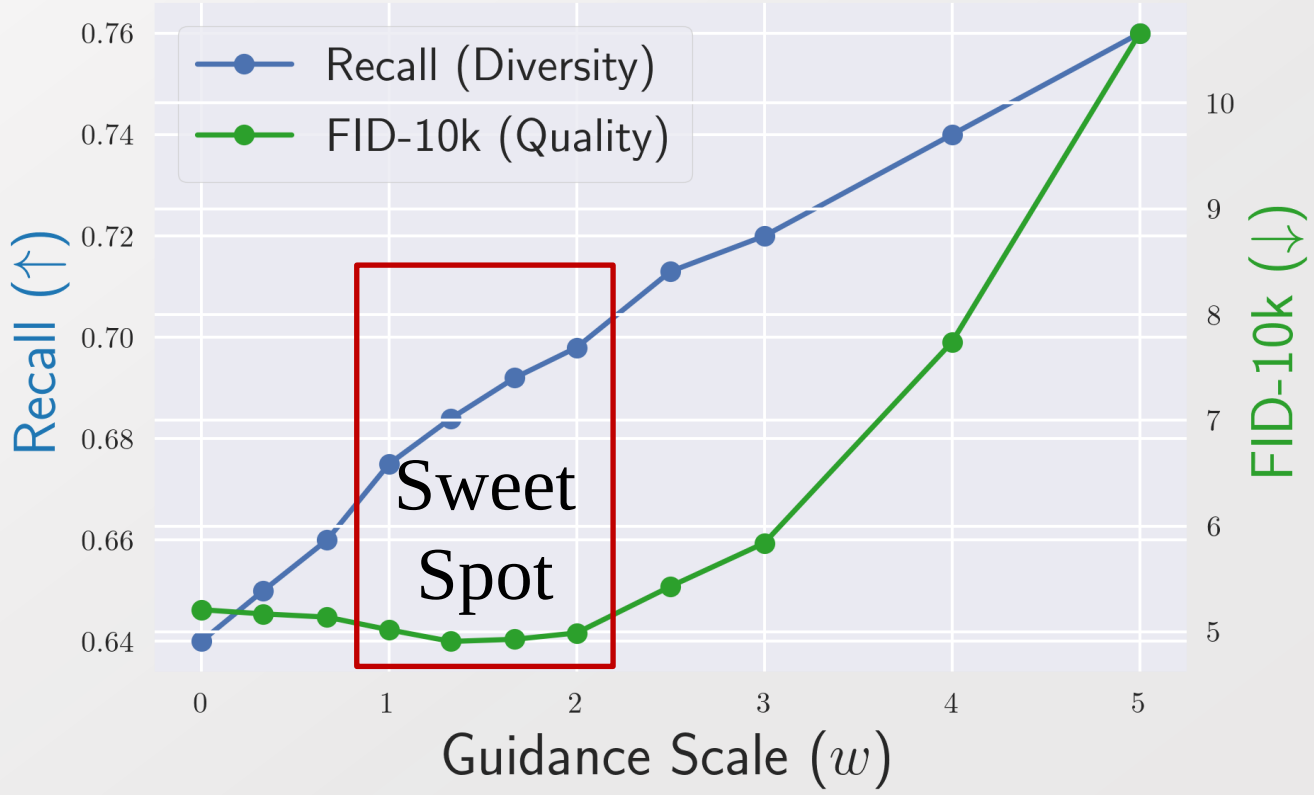
ADM
+CG+DG

Why Better Diversity?

Disc. Guidance

$$\nabla \log p^t_{model}(\mathbf{x}_t|y) + w\nabla \log \frac{d^*(\mathbf{x}_t, t|y)}{1 - d^*(\mathbf{x}_t, t|y)}$$

$$= \nabla \log \left[(p^t_{data}(\mathbf{x}_t|y))^w (p^t_{model}(\mathbf{x}_t|y))^{1-w} \right]$$



Future Study Simultaneous Training with f-Divergence

Score Mismatch Weight

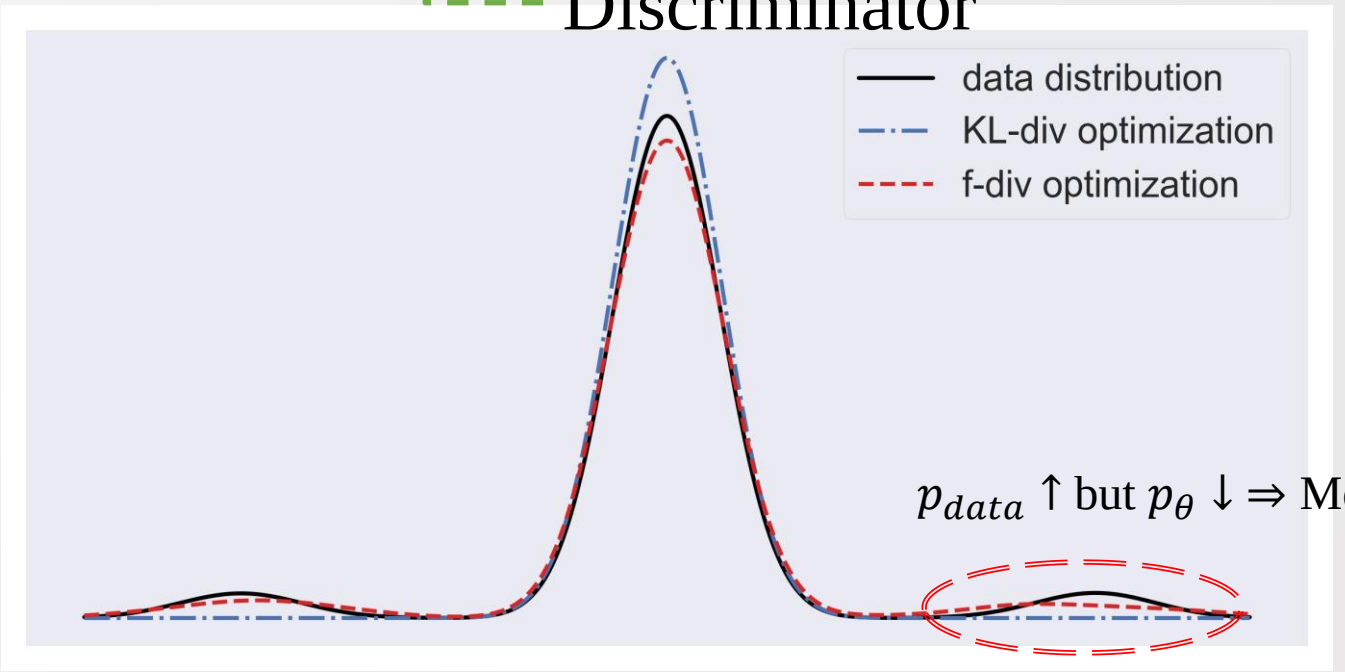
Score Mismatch

$$D_f(p_{data}(\mathbf{x}_0)\|p_{\boldsymbol{\theta}}(\mathbf{x}_0))$$

$$= \frac{1}{2}\int_0^T g^2(t)\mathbb{E}\left[f''\left(\frac{p^t_{data}(\mathbf{x}_t)}{p^t_{\boldsymbol{\theta}}(\mathbf{x}_t)}\right)\frac{p^t_{data}(\mathbf{x}_t)}{p^t_{\boldsymbol{\theta}}(\mathbf{x}_t)}\left\|\nabla \log p^t_{data}(\mathbf{x}_t) - \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)\right\|_2^2\right]\mathrm{d}t$$

$$\approx \frac{1}{2}\int_0^T g^2(t)\mathbb{E}\left[f''\left(\frac{d_{\boldsymbol{\psi}}(\mathbf{x}_t,t)}{1-d_{\boldsymbol{\psi}}(\mathbf{x}_t,t)}\right)\frac{d_{\boldsymbol{\psi}}(\mathbf{x}_t,t)}{1-d_{\boldsymbol{\psi}}(\mathbf{x}_t,t)}\left\|\nabla \log p^t_{data}(\mathbf{x}_t) - \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)\right\|_2^2\right]\mathrm{d}t$$

Discriminator



Poster
Thu 1:30 – 3:00 #804

Thank You