# Your Diffusion Model is Secretly a Noise Classifier and Benefits from Contrastive Training


Presenter: Yunshu Wu

Yunshu Wu[1] , Yingtao Luo[2], Xianghao Kong[1], Evangelos E. Papalexakis[1], Greg Ver Steeg[1]

[1] University of California Riverside

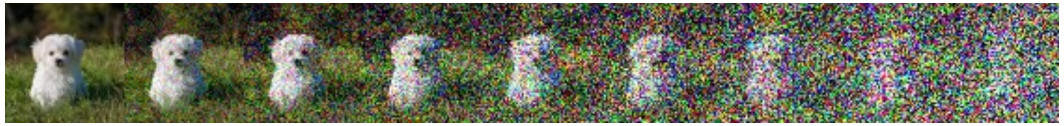
[2] Carnegie Mellon University

# Probability = (Optimal) Denoising

Guo, Shamai, Verdú, (IEEE 2005) – Information-MMSE relations
Kong, Brekelmans, Ver Steeg (ICLR 2023) ”Information-theoretic Diffusion”

Gaussian noise channel $\boldsymbol{x}_\alpha \equiv \sqrt{\sigma(\alpha)}\boldsymbol{x} + \sqrt{\sigma(-\alpha)}\boldsymbol{\epsilon}$

**log-SNR** **N(0,I)**

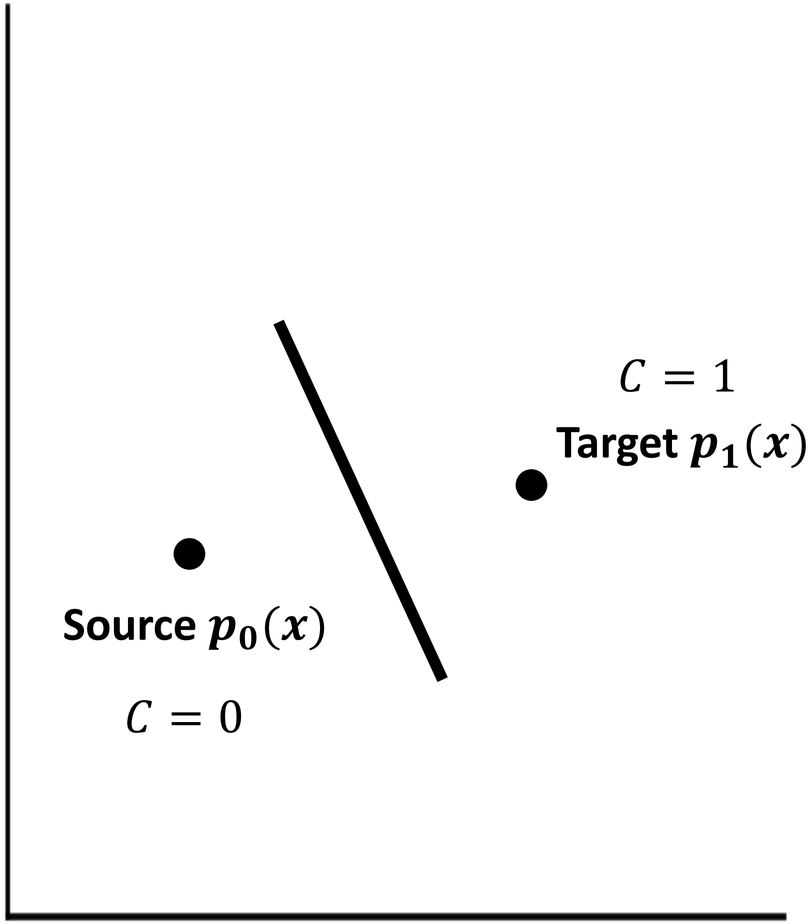



Negative Log-likelihood (NLL):

- NLL for data distribution: $$-\log p(\boldsymbol{x}) = c + {}^1\!/_2 \int_{-\infty}^{\infty} \mathbb{E}_{p(\boldsymbol{\epsilon})}[\|\boldsymbol{\epsilon} - \boxed{\hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_\alpha, \alpha)}\|_2^2]\, d\alpha.$$
- NLL for noisy data distribution: $$-\log p_\zeta(\boldsymbol{x}) = c + {}^1\!/_2 \int_{-\infty}^{\infty} d\alpha\, \mathbb{E}_{p(\boldsymbol{\epsilon})}[\|\boldsymbol{\epsilon} - \boxed{\hat{\boldsymbol{\epsilon}}_\zeta(\boldsymbol{x}_\alpha, \alpha)}\|_2^2]$$

$$-\log p_\zeta(\boldsymbol{x}) = c + {}^1\!/_2 \int_{-\infty}^{\infty} d\alpha\, \mathbb{E}_{p(\boldsymbol{\epsilon})}[\|\boldsymbol{\epsilon} - \boxed{b \cdot \hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_\alpha, \beta)}\|_2^2]$$

# Log likelihood ratios and classification




Probability space

If we sample from either distribution with probability P(C)=1/2, then,

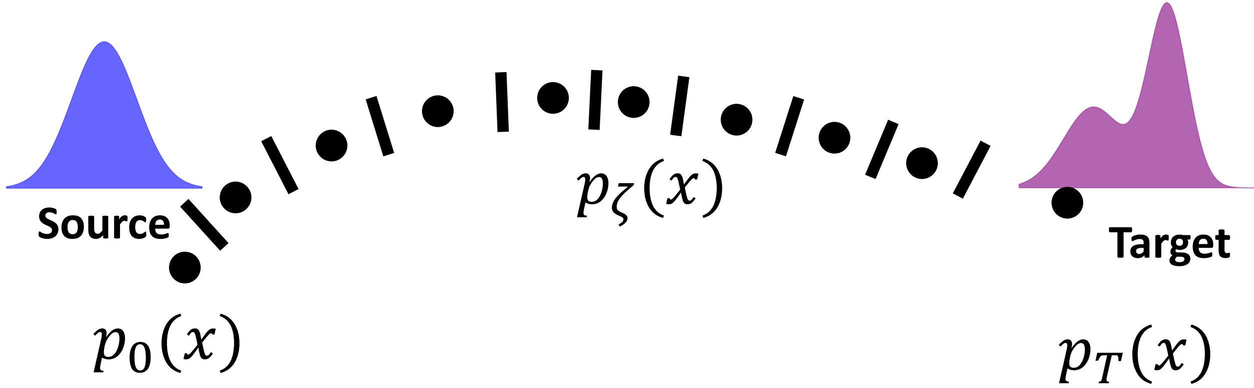
Bayes rule relates the log likelihood ratio to the optimal classifier:

$$LLR(x) \equiv \log \frac{p_1(x)}{p_0(x)} = \log \frac{\boxed{p(C = 1 \mid x)}}{\boxed{p(C = 0 \mid x)}}$$

*Idea: train classifiers to learn densities*

Gutmann & Hyvarinen (2012) “Noise contrastive estimation…”

# Denoising = Classifying




Sample from either distribution with probability q(y= $\pm$ 1) $= \frac{1}{2}$, then, the optimal binary Bayes classifier is related to the log-likelihood ratio:

$$q(y|\boldsymbol{x}) = \frac{q(\boldsymbol{x}|y)q(y)}{q(\boldsymbol{x})} \qquad \log q(y|\boldsymbol{x}) \ = -\operatorname{softplus}(y\boxed{\log \frac{q(\boldsymbol{x}|y=-1)}{q(\boldsymbol{x}|y=1)}})$$

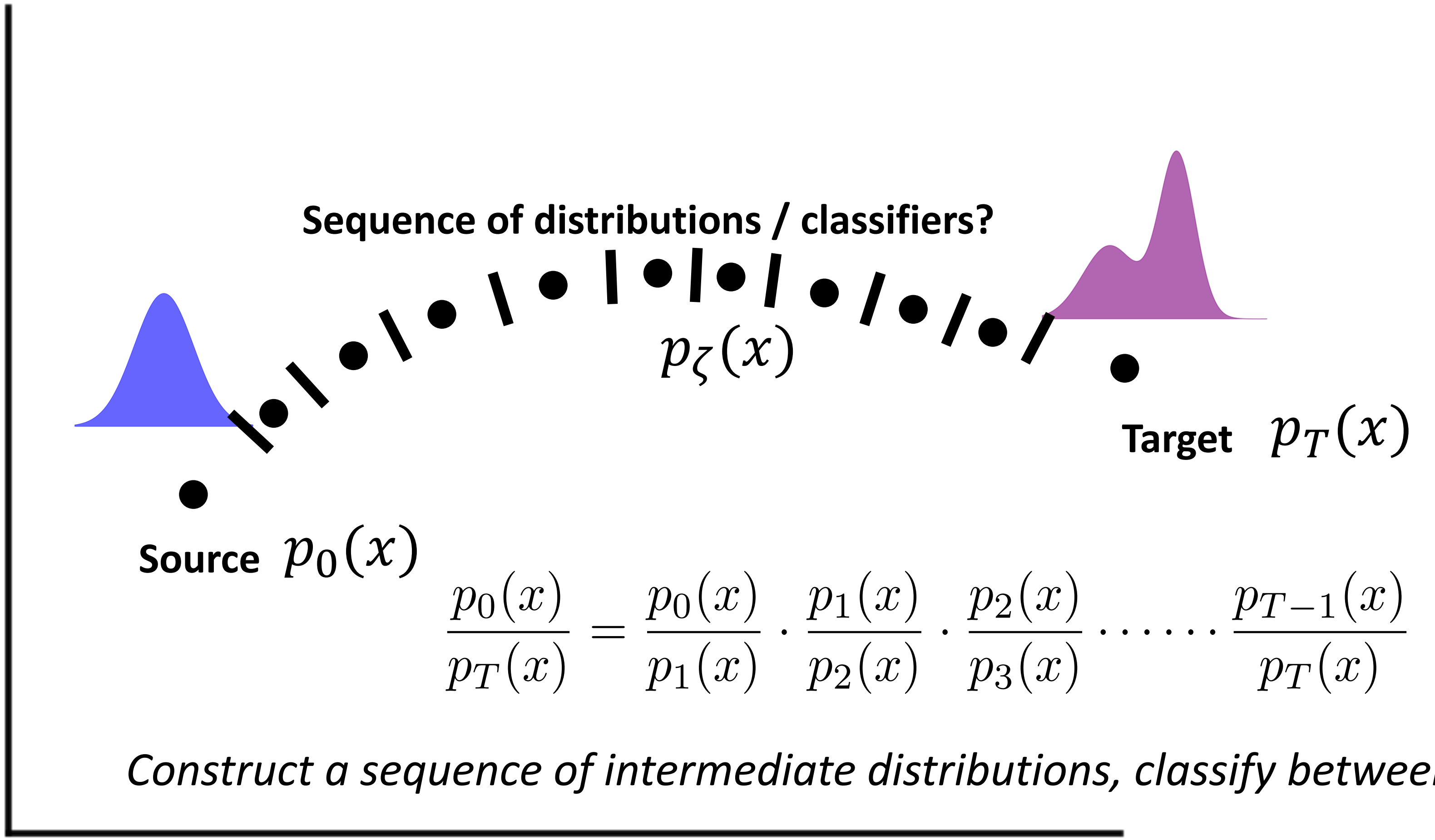
Let q(x|y=1) $= p(x)$ be data distribution, q(x|y= $-$ 1) $= p_\zeta(x)$ be noisy data distribution

LLR → Optimal classifier: $\quad LLR = \log p_\zeta(\boldsymbol{x}) - \log p(\boldsymbol{x})$

We have our Contrastive Diffusion Loss:

$$\mathcal{L}_{CDL} = \mathbb{E}_{q(\boldsymbol{x},y)}\left[\operatorname{softplus}(y(\log p_\zeta(\boldsymbol{x}) - \log p(\boldsymbol{x})))\right]$$

# CDL – more general transport schemes

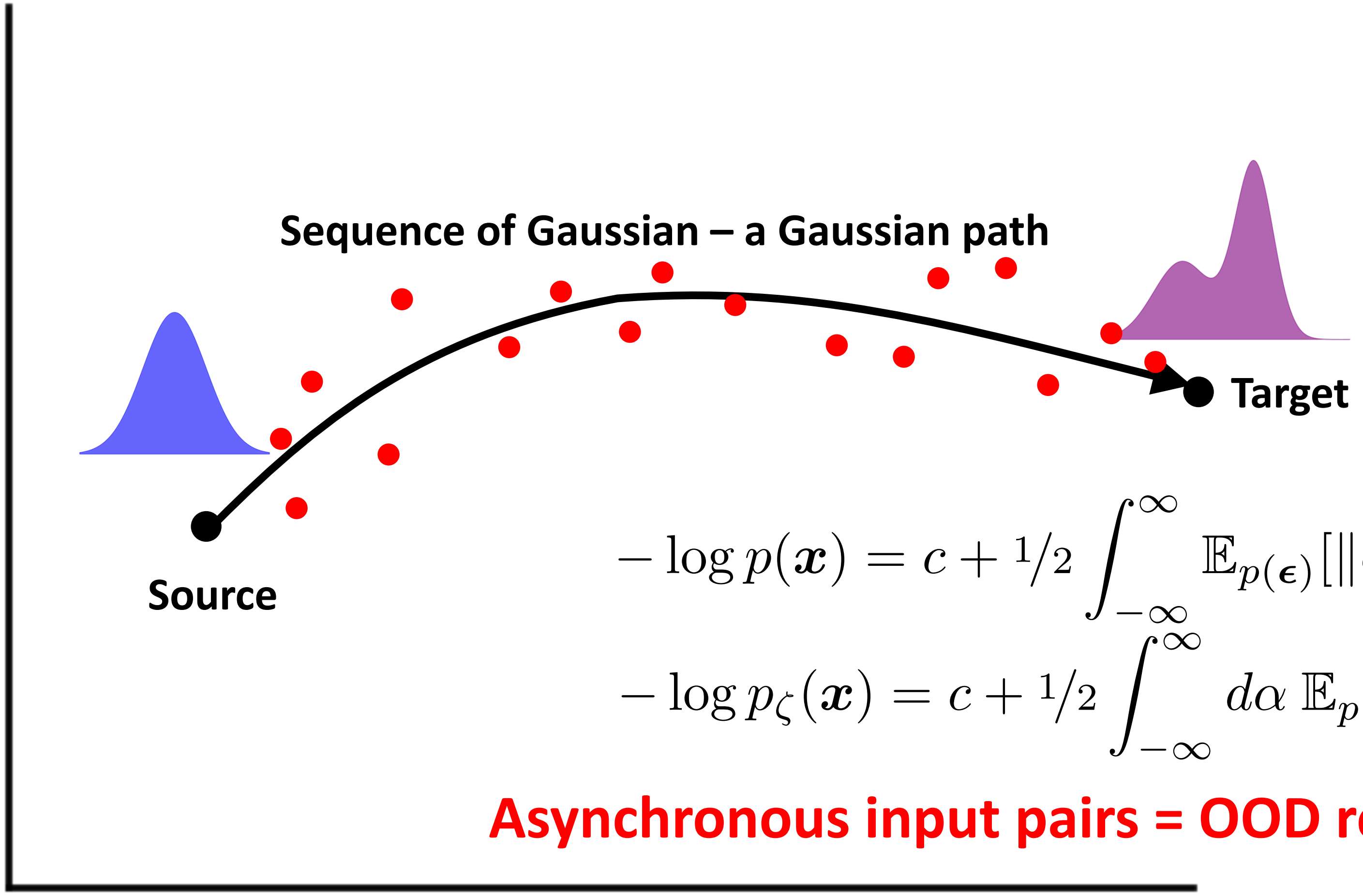



$$\frac{p_0(x)}{p_T(x)} = \frac{p_0(x)}{p_1(x)} \cdot \frac{p_1(x)}{p_2(x)} \cdot \frac{p_2(x)}{p_3(x)} \cdots\cdots \frac{p_{T-1}(x)}{p_T(x)}$$

*Construct a sequence of intermediate distributions, classify between them.*

Rhodes et al (NeurIPS 2022) “Telescoping density-ratio estimation”

# Profit : CDL training on OOD regions

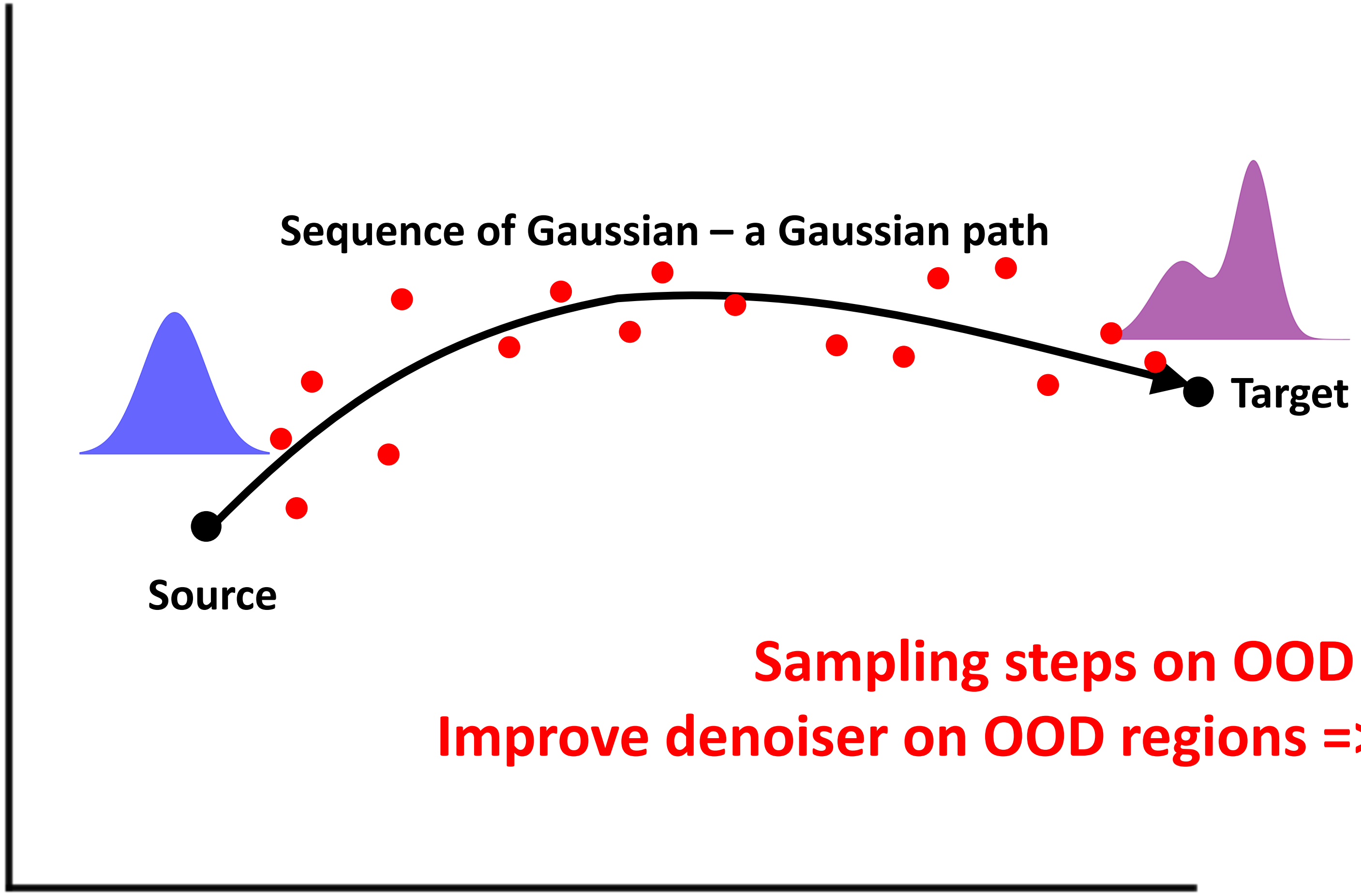



$$-\log p(\boldsymbol{x}) = c + {}^{1}\!/_{2} \int_{-\infty}^{\infty} \mathbb{E}_{p(\boldsymbol{\epsilon})}[\|\boldsymbol{\epsilon} - \boxed{\hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_\alpha, \alpha)}\|_2^2]\, d\alpha.$$

$$-\log p_\zeta(\boldsymbol{x}) = c + {}^{1}\!/_{2} \int_{-\infty}^{\infty} d\alpha\, \mathbb{E}_{p(\boldsymbol{\epsilon})}[\|\boldsymbol{\epsilon} - \boxed{b \cdot \hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_\alpha, \beta)}\|_2^2]$$

**Asynchronous input pairs = OOD regions for std diffusion loss**

# Profit : CDL training on OOD regions

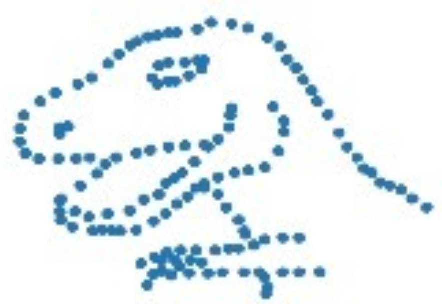



**Sampling steps on OOD regions**

**Improve denoiser on OOD regions => improve sampling**

# Better distribution learning with hard constraints

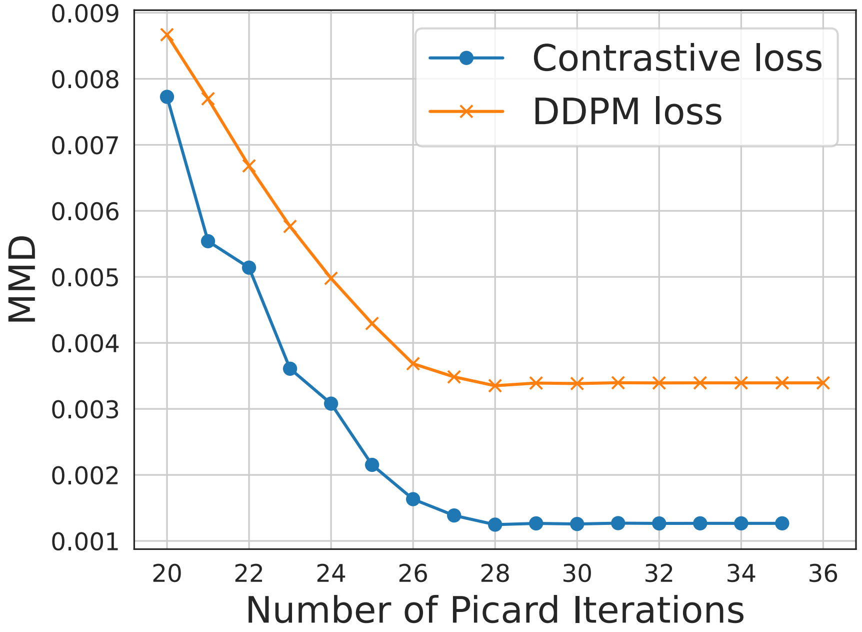
Data



Contrastive

Standard diffusion

Distribution similarity




Compute

# Contrastive training improves “FID”

| Models | CIFAR-10 at 32x32 | | AFHQv2 64x64 | FFHQ 64x64 |
|---|---|---|---|---|
| | unconditional | conditional | unconditional | unconditional |
| DDPM | 9.43 | NA | NA | NA |
| CDL-DDPM | **9.06** | NA | NA | NA |
| VP | $3.24 \pm 0.02$ | $2.93 \pm 0.02$ | $2.95 \pm 0.03$ | $3.67 \pm 0.04$ |
| CDL-VP | $\mathbf{2.51 \pm 0.01}$ | $\mathbf{2.41 \pm 0.01}$ | $\mathbf{2.91 \pm 0.02}$ | $\mathbf{3.33 \pm 0.03}$ |
| VE | $3.00 \pm 0.01$ | $2.76 \pm 0.01$ | $2.98 \pm 0.03$ | $3.65 \pm 0.02$ |
| CDL-VE | $\mathbf{2.38 \pm 0.01}$ | $\mathbf{2.25 \pm 0.02}$ | $\mathbf{2.93 \pm 0.01}$ | $\mathbf{3.29 \pm 0.02}$ |

Table 2: Evaluating FID score (lower is better) of parallel DDPM sampler on real-world datasets using $5,0000$ samples. “NA” stands for "Not Applicable". For reported FID scores, we run three sets of random seeds and reported the average with uncertainty.

# Thank you!