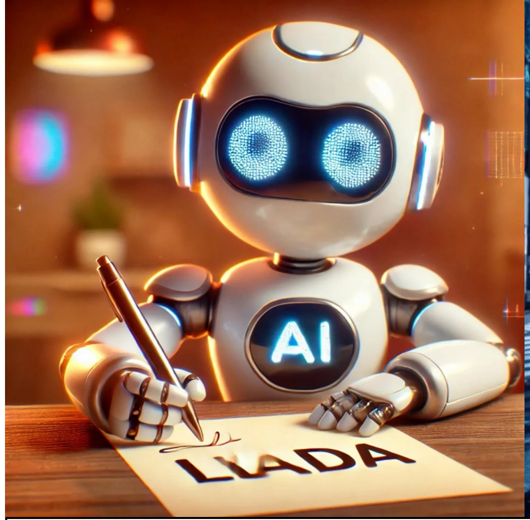
Learning Process & Sampling Complexity of Diffusion Models

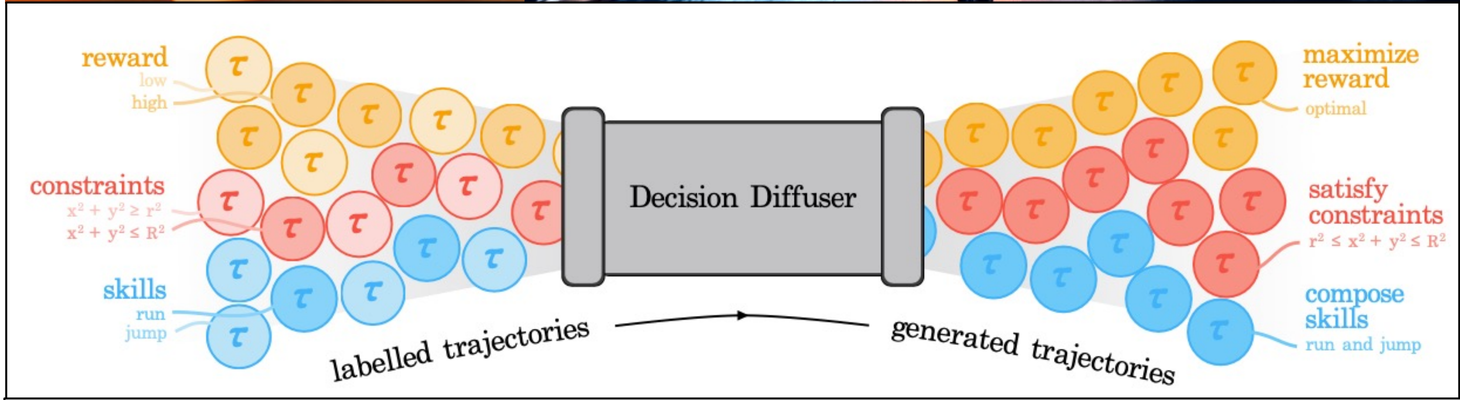
Shuai Li

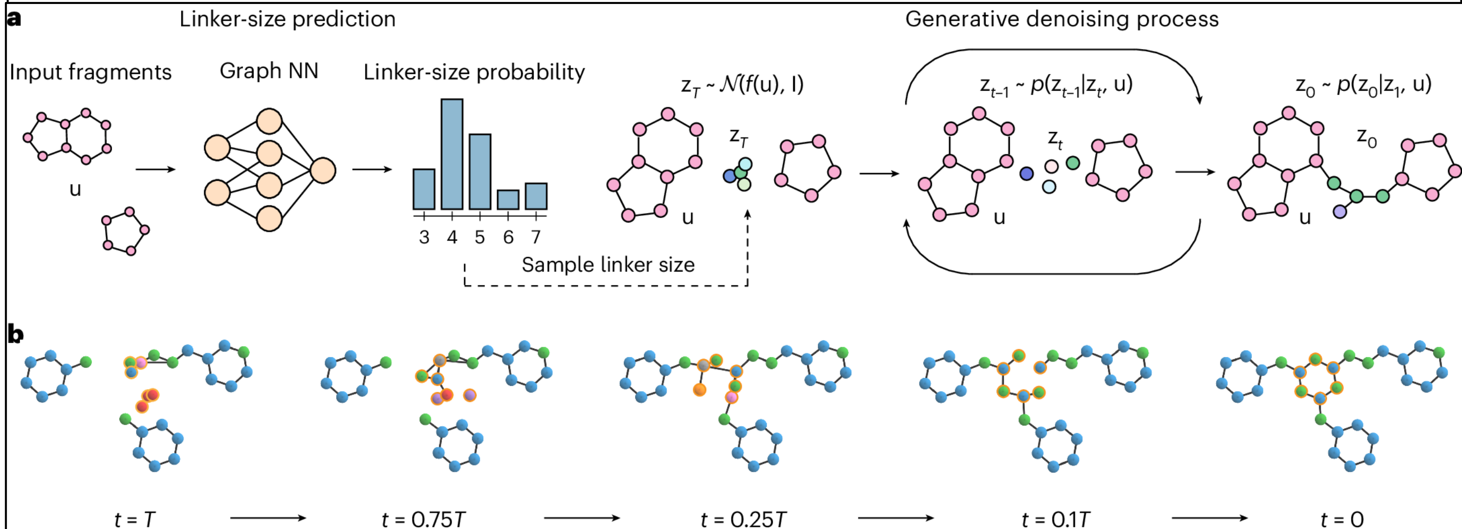
2025.11

Diffusion Models

- Vision: Sora, etc.
    - SOTA result: Image, 3D, video
- Language: LLaDA
- Multi-modal Models: MMaDA
- Reinforcement Learning
- AI4Science







[1] NZYZOHZLWL, Large Language Diffusion Models, ICLR 2025 DeLTa Workshop, Oral.
[2] YTLZSTW, Multimodal Large Diffusion Language Models, NeurIPS 2025.
[3] ADGTJA, Is Conditional Generative Modeling all you need for Decision Making?, ICLR 2023.
[4] ISVSSFWBC, Equivariant 3D-conditional diffusion model for molecular linker design, Nature Machine Intelligence 2024.

Theory Helps Training & Sampling

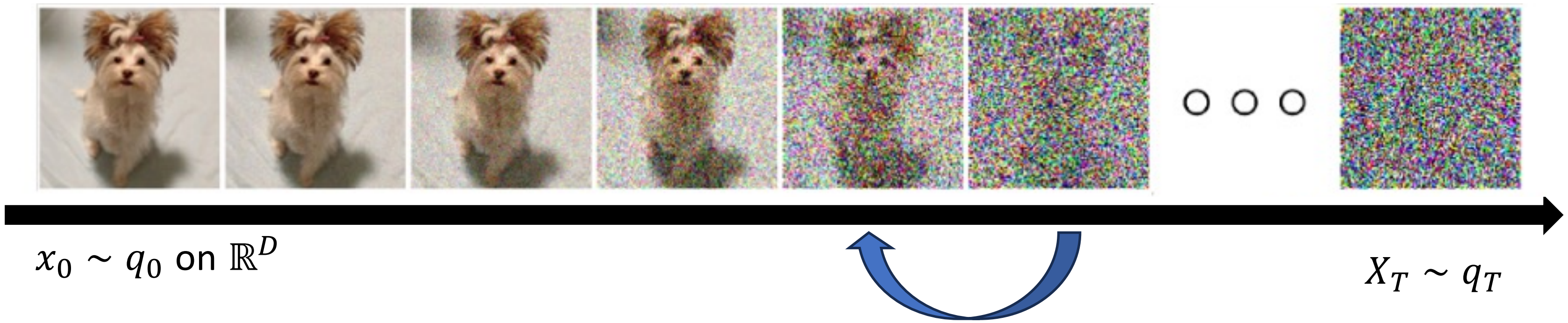
- Solid theoretical foundation helps efficient training & fast sampling:

- Theoretical SDE framework of diffusion family unifies training & sampling[1]
- New training paradigm with SOTA performance: Flow-matching[2]
- $10\times$ Faster sampling algorithm: DPM-Solvers series[3], Analytic-DPM[4]

[1] SDKKEP, Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021.
[2] LG, Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, ICLR 2023.
[3] LZBCLZ, Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, NeurIPS 2022.
[4] BLZZ, Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models, ICLR 2022.

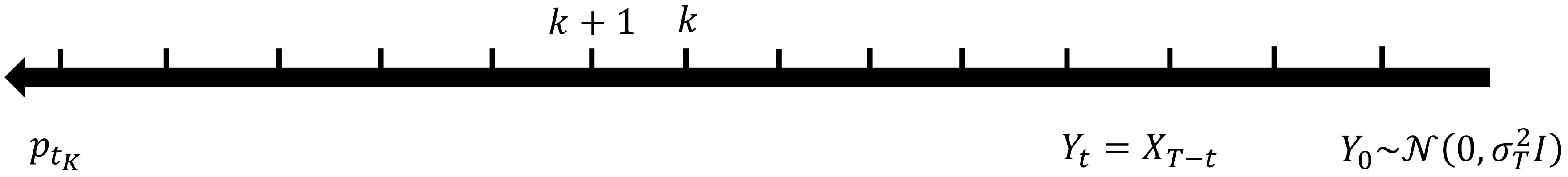
Paradigm of Multi-step Diffusion Models

Forward Process



Core Problem 1: Training Process to Learn Denoising

Reverse Process

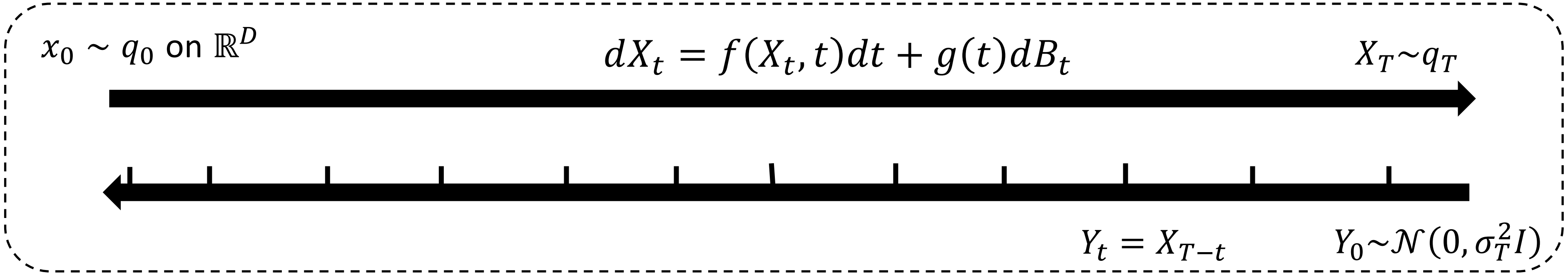


Core Problem 2: Sampling Complexity K

Overview

- Pretraining: Efficient Multi-manifold MoG Model
- Fine-tuning: Good Sharing Latent Guarantees Few-shot Efficiency
- Sampling: Complexity for Multi-step Diffusion Models
- Discretization: Complexity of 1-step Models in Training Phase

Mathematical Framework of Diffusion Models



score function

core to train DM unknown

- $dY_t = \left[-f(Y_t, T-t) + \frac{1+\eta^2}{2} g^2(T-t) \nabla \log q_{T-t}(Y_t) \right] dt + \eta g(T-t) dB_t, \eta \in [0,1]$

- Score matching training objective:

conditional distribution known

$$\min_{s \in \mathfrak{F}} \hat{\mathcal{L}}(s) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T-\delta} \int_{\delta}^{T} \mathbb{E}_{X_t | X_0 = X_i} \left[\| \nabla \log q_t(X_t | X_0) - s(X_t, t) \|_2^2 \right] dt$$

Learning Faces Curse of Dimension

- Minimiser $s_\theta \in \operatorname{argmin}_\Theta \hat{\mathcal{L}}(s)$ satisfies

  Estimation Error= $\frac{1}{T-\delta}\int_\delta^T \mathbb{E}_{q_t}\left[\|\nabla \log q_t(X_t) - s_\theta(X_t,t)\|_2^2\right]dt < O\left(n^{-1/D}\right)$

  covering number & concentration

  $D = 3\times256\times256 \approx 2\times10^5$

- Good training requires training data size $n = O(10^{10^5})$ Huge!!

- Efficient training needs utilizing data structure!

Data Structures: Existing Works

Manifold Modeling	Latent		# of Parameters	Estimation Error
Full Space [1]	General	X	$O(D^{D+1})$	$O(n^{-1/D})$
Full Space [2]	Mixture of Gaussian (MoG)	$X \sim \sum_{m=1}^{M} \pi_m \mathcal{N}(\mu_m, \Sigma_m)$	$O(MD^2)$	$O(\frac{\sqrt{DM}}{\sqrt{n}})$
Low-dim manifold [3]	General	$X = Az$, with $A \in \mathbb{R}^{D \times d}$	$O(Dd + d^{d+1})$	$O(n^{-\frac{2}{d}})$
Multi-manifold	General	$X = \sum_{\ell=1}^{L} \pi_\ell A_\ell z_\ell$, with $A_\ell \in \mathbb{R}^{D \times d}$	$O(LDd + Ld^{d+1})$	$O(\sqrt{L}n^{-\frac{2}{d}})$
Multi-manifold [4]	Gaussian	$X \sim \sum_{\ell=1}^{L} \pi_\ell \mathcal{N}(\cdot; 0, A_\ell A_\ell^\top)$	$O(LDd)$	$O(\frac{\sqrt{dL}}{\sqrt{n}} + \text{Const})$

[1] OAS, Diffusion Models are Minimax Optimal Distribution Estimators, ICML 2023.

[2] SCK, Learning mixtures of gaussians using the ddpm objective, NeurIPS 2023.

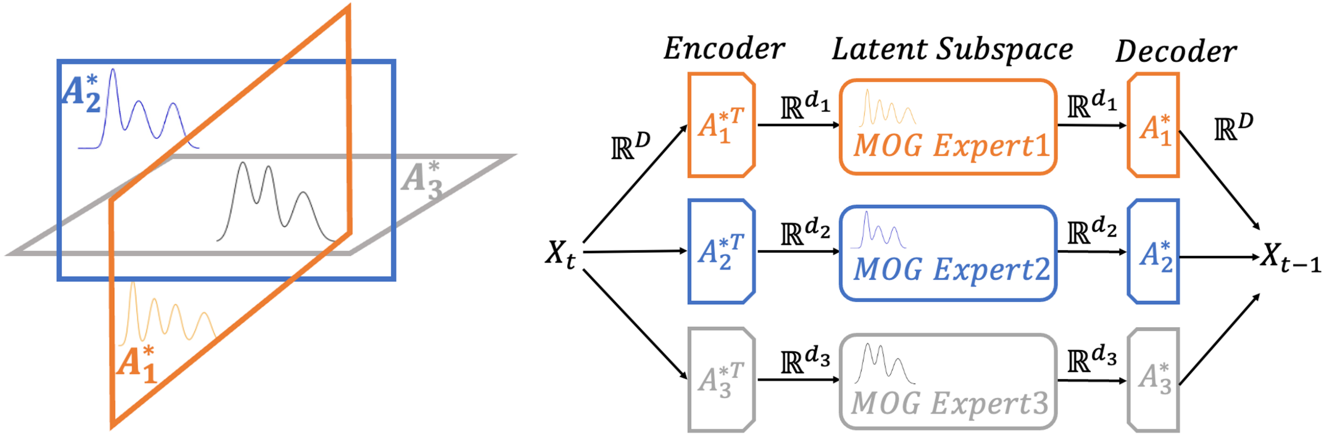
[3] CHZW, Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data, ICML 2023.

[4] WZZCMQ. Diffusion models learn low-dimensional distributions via subspace clustering, NeurIPS 2024 M3L Workshop.

Multi-manifold Mixture-of-Gaussian Modeling

- $X \sim \sum_{\ell=1}^{L} \pi_\ell \sum_{m=1}^{M} \pi_{\ell,m} \mathcal{N}\left(\cdot; A_\ell \mu_{\ell,m}, A_\ell \Sigma_{\ell,m} A_\ell^\top\right)$ Most general!
- Theorem. Its estimation error satisfies

$$\frac{1}{T-\delta}\int_{\delta}^{T} \mathbb{E}_{q_t}\left[\|\nabla \log q_t(X_t) - s_\theta(X_t,t)\|_2^2\right] dt < O\left(\frac{\sqrt{LM}\sqrt{dL}}{\sqrt{n}}\right)$$

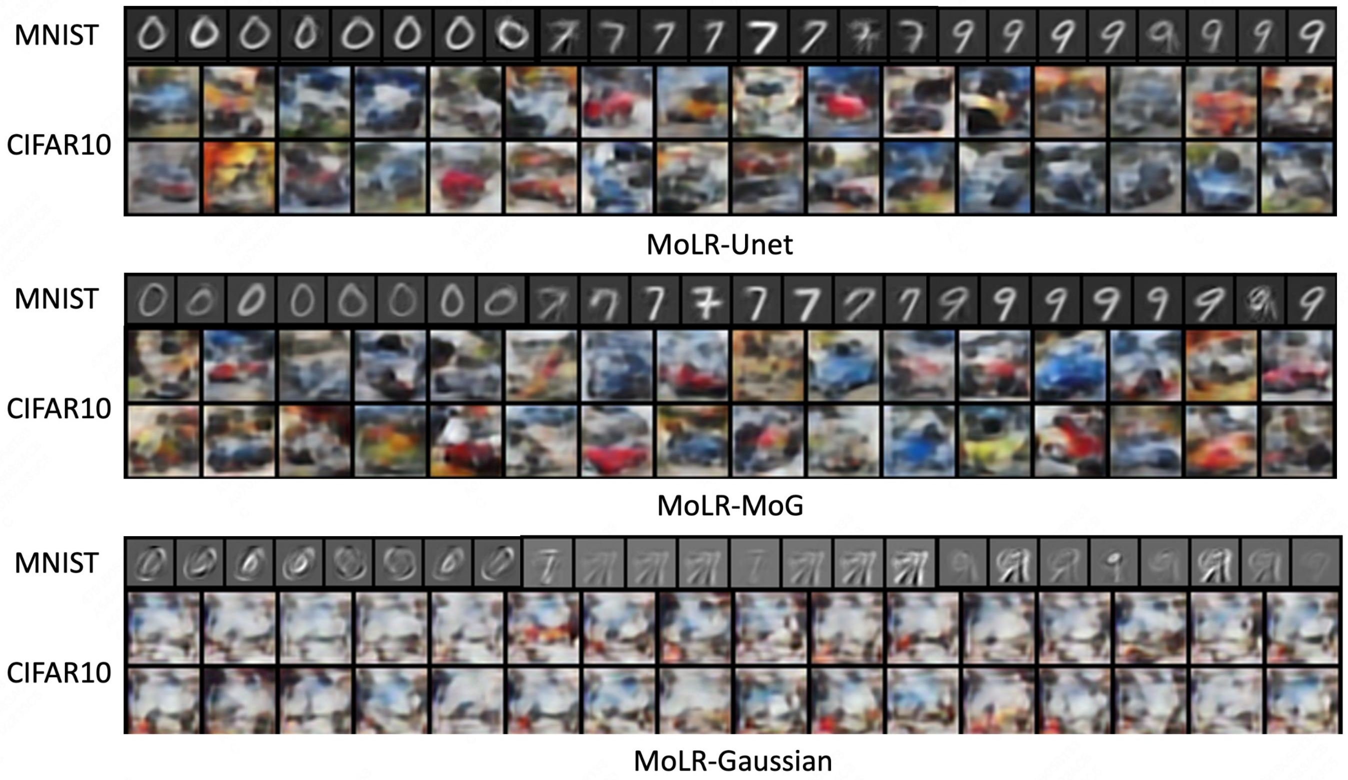


(a) MoLR-MoG Modeling **(b) MoE-nonlinear MoG Score**

[1] YLJCL, Multi-Subspace Multi-Modal Modeling for Diffusion Models: Estimation, Convergence and Mixture of Experts (In submission)

Much Smaller Model w/ Sufficiently Good Performance

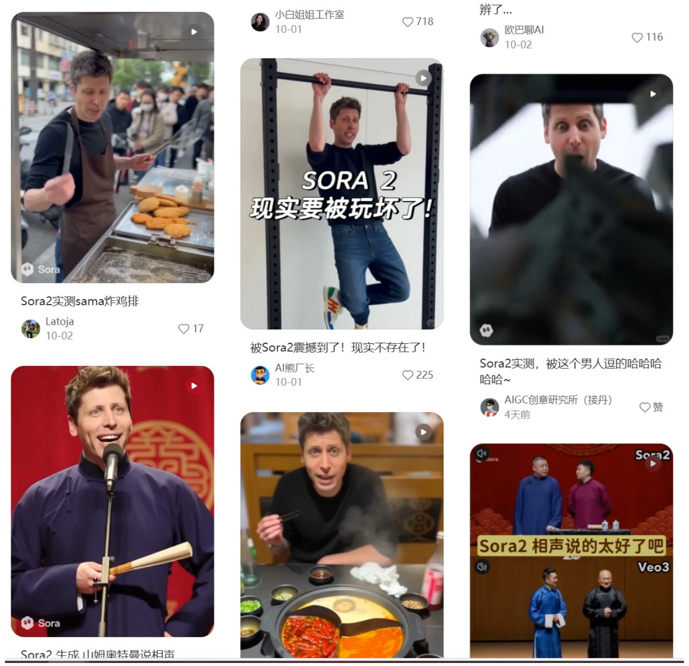
Latent	# of Parameters	Estimation Error	MNIST Acc/ Performance
General	$O(LDd + Ld^{d+1})$	$O(\sqrt{L}n^{-\frac{2}{d}})$	0.96 √ Deep NN
Mixture of Gaussian	$O(LDd + Ld^2)$	$O\left(\frac{\sqrt{LM}\sqrt{dL}}{\sqrt{n}}\right)$	0.89 √ 2-layer NN
Gaussian	$O(LDd)$	$O(\frac{\sqrt{dL}}{\sqrt{n}} + \text{Const})$	0.08 × Linear NN



- YLJC**L**, Multi-Subspace Multi-Modal Modeling for Diffusion Models: Estimation, Convergence and Mixture of Experts (In submission)

Overview

- Pretraining: Efficient Multi-manifold MoG Model
- Fine-tuning: Good Sharing Latent Guarantees Few-shot Efficiency
- Sampling: Complexity for Multi-step Diffusion Models
- Discretization: Complexity of 1-step Models in Training Phase



Few-shot Fine-tuning is key to the customized creation

but no theory supports effective information sharing

Few-shot Fine-tuning

- Pretrain w/ large source data (2.3 Billion): $\{X_{s,i}\}_{i=1}^{n_s} \sim q_0^s$ on $\mathbb{R}^D$
- $\min_{s \in \text{Source } \mathfrak{P}} \hat{\mathcal{L}}_s(s) = \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{1}{T-\delta} \int_\delta^T \mathbb{E}_{X_t|X_0=X_{s,i}} \left[\|\nabla \log q_t^s(X_t|X_0) - s(X_t,t)\|_2^2 \right] dt$
- Estimation error $O(n_s^{-\frac{2}{d}})$

e.g. 882M

Tolerable!

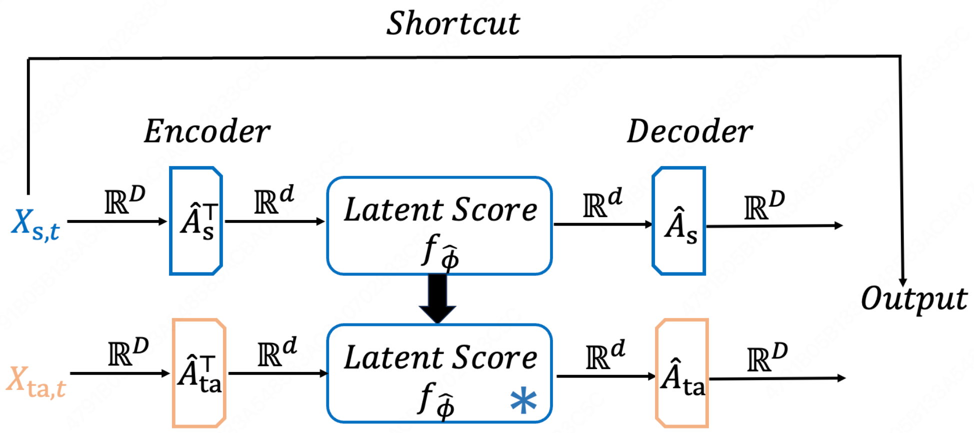
- Fine-tune with limited target data (~10 images): $\{X_{ta,i}\}_{i=1}^{n_{ta}} \sim q_0^{ta}$
- $\min_{s \in \text{Target } \mathfrak{P}} \hat{\mathcal{L}}_{ta}(s) = \frac{1}{n_{ta}} \sum_{i=1}^{n_{ta}} \frac{1}{T-\delta} \int_\delta^T \mathbb{E}_{X_t|X_0=X_{ta,i}} \left[\|\nabla \log q_t^{ta}(X_t|X_0) - s(X_t,t)\|_2^2 \right] dt$
- Estimation error $O(n_{ta}^{-\frac{2}{d}})$

e.g. 1.5M 0.17%

Meaningless!

Information-sharing Model Design

- Empirical works share most parameters and fine-tune key parameters
- Assumption. The source and target data admit linear structure and share latent space $X_s = A_s z$ and $X_{ta} = A_{ta} z, z \in \mathbb{R}^d$

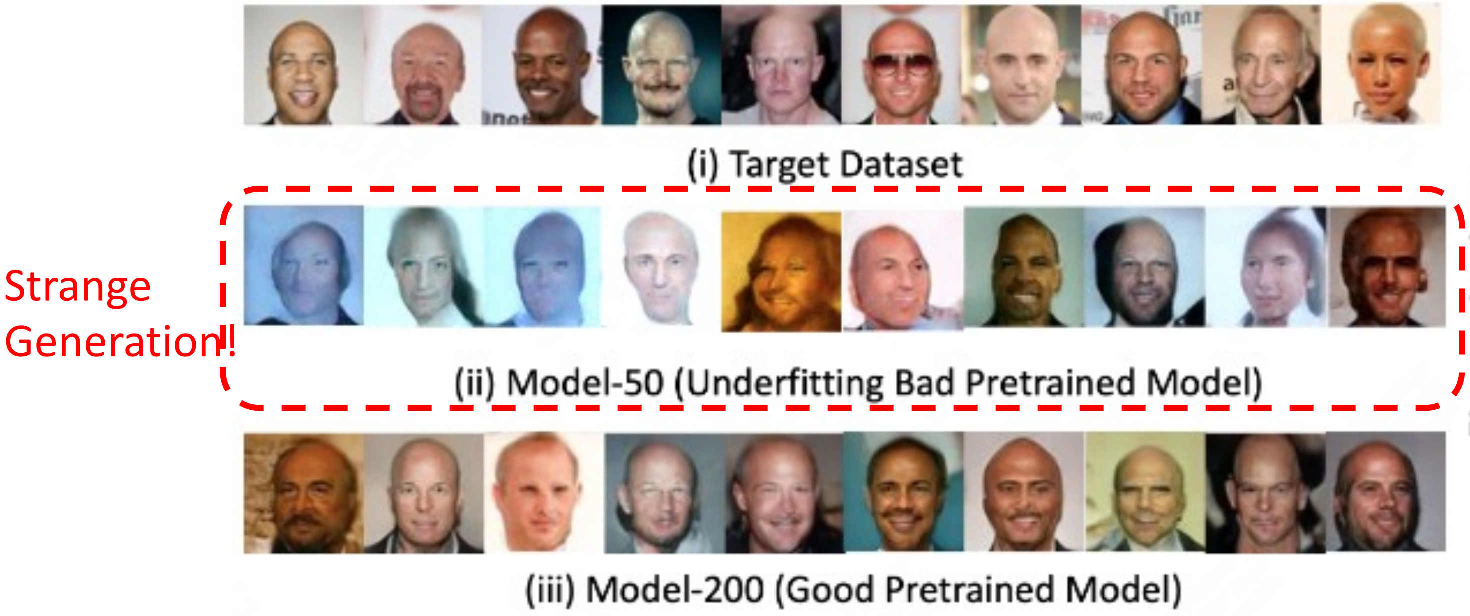


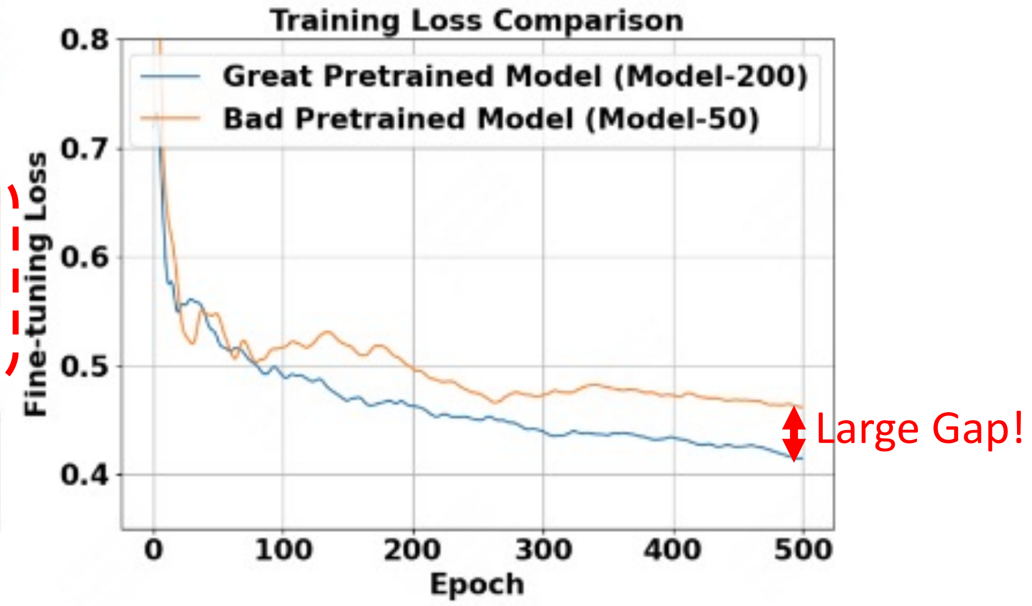
- Then the score function is

$$\nabla \log q_t^{ta}(X) = A_{ta} \nabla \log q_t^{\text{Latent}}(A_{ta}^\top X) - \frac{1}{\sigma_t^2}(I_D - A_{ta} A_{ta}^\top) X$$

Shared Latent Score

Bad Latent Leads to Large Estimation Error





- Theorem. W/ bad latent

$$\frac{1}{T-\delta}\int_{\delta}^{T} \mathbb{E}_{q_t^{\mathrm{ta}}}\left[\|\nabla \log q_t^{\mathrm{ta}}(X_t) - s_\theta(X_t, t)\|_2^2\right] dt \geq \mathrm{Const}$$

[1] YLJCL, Evaluating the Role of Great Pre-trained Diffusion Models in Few-shot Phase: Warm-up and Acceleration (In submission).

Bad Latent Suffers Bad Local Minima

Prompt cat but results in dog figure

Bad latent fails to fit target feature!

- Theorem. W/ bad latent, $\exists\, s_{\theta}^{\text{few-shot}} \neq s_{\theta^*}^{\text{few-shot}}$ s.t. $\frac{\partial s_{\theta}^{\text{few-shot}}}{\partial \theta} \approx 0$

[1] YLJCL, Evaluating the Role of Great Pre-trained Diffusion Models in Few-shot Phase: Warm-up and Acceleration (In submission)

Good Latent Secures Efficiency

- Theorem. The estimation error of few-shot diffusion model is

$$\frac{1}{T-\delta}\int_{\delta}^{T}\mathbb{E}_{q_t^{\mathrm{ta}}}\left[\left\|\nabla\log q_t^{\mathrm{ta}}(X_t)-s_{\hat{A}_{\mathrm{ta}},\hat{\phi}}(X_t,t)\right\|_2^2\right]dt\leq O\left(n_{\mathrm{ta}}^{-\frac{1}{2}}+n_s^{-\frac{2}{d}}\right)$$

Guarantee good latent

- $O\left(n_{\mathrm{ta}}^{-\frac{1}{2}}\right)$ explains why $5-8$ images are enough for few-shot fine-tuning

Table 1: The requirement of n_{ta} in popular datasets. We use latent dimension in Pope et al. (2021).

Dataset	CIFAR-10	CIFAR-100	CelebA	MS-COCO	ImageNet
Dataset Size	6×10^4	6×10^4	2×10^5	3.3×10^5	1.2×10^6
Latent Dimension	25	22	24	37	43
The Requirement of n_{ta}	6	8	8	5	5

- YHCJWL, Few-shot diffusion models escape the curse of dimensionality. NeurIPS 2024.

Good Latent Leads to Good Landscape

- Theorem. With a good shared latent, the landscape of the few-shot optimization is κ-strongly convex w/ convergence rate

$$\left\| \hat{A}_{\mathrm{ta}}^{(i)} \hat{A}_{\mathrm{ta}}^{(i)\top} - A_{\mathrm{ta}} A_{\mathrm{ta}}^{\top} \right\|_F \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^i \|A_{\mathrm{ta}}\|_F \left\| \hat{A}_{\mathrm{ta}}^{(0)} - A_{\mathrm{ta}} \right\|_F$$

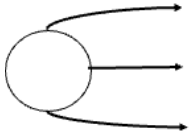
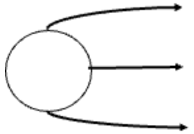
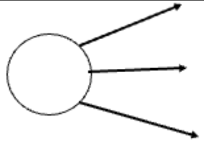
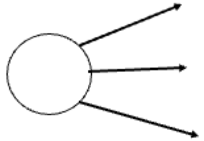
[1] YLJCL, Evaluating the Role of Great Pre-trained Diffusion Models in Few-shot Phase: Warm-up and Acceleration (In submission)

Overview

- Pretraining: Efficient Multi-manifold MoG Model
- Fine-tuning: Good Sharing Latent Guarantees Few-shot Efficiency
- Sampling: Complexity for Multi-step Diffusion Models
- Discretization: Complexity of 1-step Models in Training Phase

Common Forward Processes

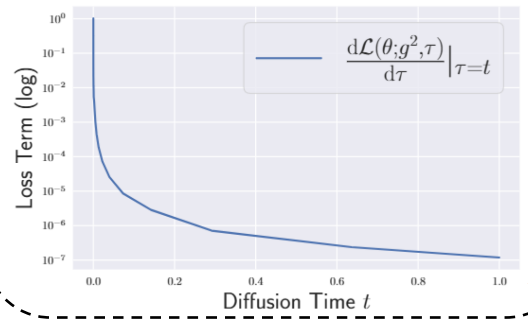
$$dX_t = f(X_t, t)dt + g(t)dB_t \longrightarrow T$$

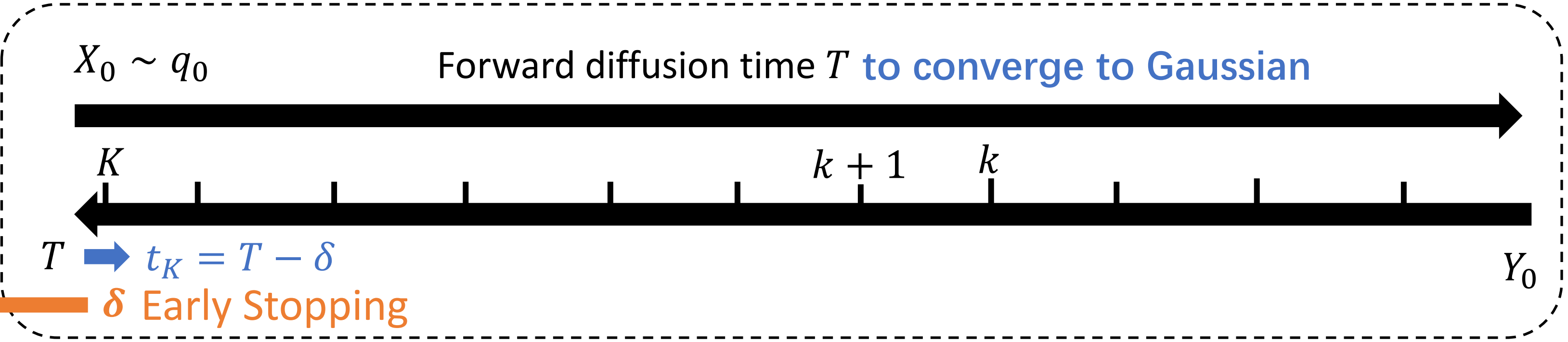
		Trajectory	Forward Distribution	
Variance Preserving (VP) [1]	$f(X_t, t) = -\frac{1}{2}X_t$ $g(t) = 1$		$\mathcal{N}(0, I_D)$	stability.ai, Midjourney
Variance Exploding (VE–SMLD) [2]	$f(X_t, t) = 0$ $g(t) = \sqrt{2}$		$\mathcal{N}(0, TI_D)$	Stanford University
Variance Exploding (VE–EDM) [3]	$f(X_t, t) = 0$ $g(t) = \sqrt{2t}$		$\mathcal{N}(0, T^2 I_D)$	
Rectified Flow (RF) [4]	$X_t = (1-t)X_0 + tZ$ $t \in [0,1]$		$\mathcal{N}(0, I_D)$	

[1] HJA, Denoising diffusion probabilistic models, NeurIPS 2020.
[2] SE, Generative modeling by estimating gradients of the data distribution, NeurIPS 2019.
[3] KAAL, Elucidating the Design Space of Diffusion-Based Generative Models, NeurIPS 2022.
[4] LG, Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, ICLR 2023.

Sampling Complexity: Objective

to avoid Score blow-up





- Objective:

  With accurate score $\|\nabla \log q_t(X) - s_\theta(X,t)\|_2^2 \leq \epsilon_{\text{score}}^2$

  Minimize sample complexity K s.t.

$$\mathrm{KL}\left(p_{t_K}, q_\delta\right) \leq \epsilon_{\mathrm{KL}}^2 \text{ and } W_2^2(q_0, q_\delta) \leq \epsilon_{W_2}^2$$

Sample Complexity: General Guarantee for Reverse SDE

- Theorem. Sample complexity can be divided by

Convergence of Forward Process

Discretization

$$\mathrm{KL}\left(p_{t_K}, q_\delta\right) \le \mathrm{KL}(\mathcal{N}(0,\sigma_T^2), q_T) + \sum_{k=0}^{K-1} \mathbb{E}_{q_{t_k}(x)} \mathrm{KL}\left(p_{t_{k+1}|t_k}(\cdot\,|x), q_{t_{k+1}|t_k}(\cdot\,|x)\right)$$

$$\le D^2 m_T/\sigma_T^2 + D^2 (T/\delta)^{\frac{1}{a}}/K \le \tilde{O}\left(\epsilon_{\mathrm{KL}}^2\right)$$

- Then the sample complexity requires $K = O(D^2 (T/\delta)^{\frac{1}{a}}/\epsilon_{\mathrm{KL}}^2)$ where δ satisfies

$$\mathrm{W}_2^2(q_0, q_\delta) \le \sigma_\delta^2 \le \epsilon_{W_2}^2$$

Sample Complexities

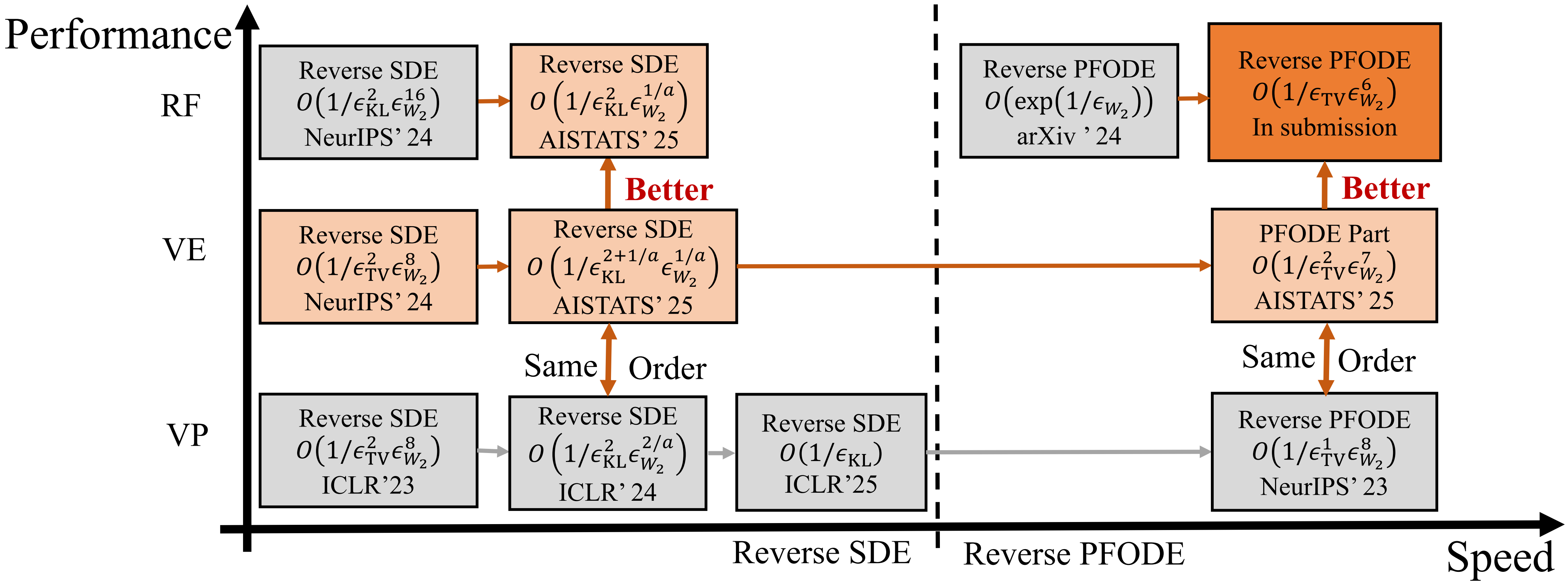
	m_T	σ_T^2	$\boldsymbol{T}$: $\mathrm{KL}(\mathcal{N}(0,\sigma_T^2), q_T) \le \frac{m_T}{\sigma_T^2} \le \epsilon_{\mathrm{KL}}^2$	σ_δ^2	$\boldsymbol{\delta}$: $\mathrm{W}_2^2(q_0, q_\delta) \le \sigma_\delta^2 \le \epsilon_{W_2}^2$	$\boldsymbol{K}$: $O(D^2(T/\delta)^{\frac{1}{a}}/\epsilon_{\mathrm{KL}}^2)$
VP	e^{-T}	$1 - e^{-2T}$	$\log(1/\epsilon_{\mathrm{KL}})$ √	δ	$\epsilon_{W_2}^2$ ×	$O\left(D^2/\epsilon_{\mathrm{KL}}^2 \epsilon_{W_2}^{2/a}\right)$
VE (SMLD)	1	T	$1/\epsilon_{\mathrm{KL}}^2$ ×	δ	$\epsilon_{W_2}^2$ ×	$O\left(D^2/\epsilon_{\mathrm{KL}}^{2+2/a} \epsilon_{W_2}^{2/a}\right)$
VE (EDM)	1	T^2	$1/\epsilon_{\mathrm{KL}}$ ×	δ^2	ϵ_{W_2} √	$O\left(D^2/\epsilon_{\mathrm{KL}}^{2+1/a} \epsilon_{W_2}^{1/a}\right)$
RF	1	1	1 √	δ^2	ϵ_{W_2} √	$O\left(D^2/\epsilon_{\mathrm{KL}}^2 \epsilon_{W_2}^{1/a}\right)$

- VP better in T and VE (EDM) better in δ
- RF better in both T and δ and thus has a better complexity

Results Extend to PRODE

[1] YWJL, Leveraging drift to improve sample complexity of variance exploding diffusion models. NeurIPS 2024.
[2] YJL, The Polynomial Iteration Complexity for Variance Exploding Diffusion Models: Elucidating SDE and ODE Samplers. AISTATS 2025.
[3] YZJCL, Elucidating Rectified Flow with Deterministic Sampler: Polynomial Discretization Complexity for Multi and One-step Models. Arxiv.

- Reverse SDE generate diverse samples while PFODE generate fast



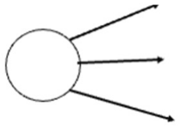
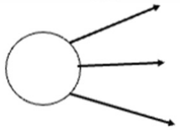
Overview

- Pretraining: Efficient Multi-manifold MoG Model
- Fine-tuning: Good Sharing Latent Guarantees Few-shot Efficiency
- Sampling: Complexity for Multi-step Diffusion Models
- Discretization: Complexity of 1-step Models in Training Phase

Linear Trajectory & PFODE Achieve 1-step Generation

- PFODE generate deterministically compared to reverse SDE

- VE-EDM and RF have linear trajectory

Variance Exploding (VE–EDM) [3]	$f(X_t, t) = 0$ $g(t) = \sqrt{2t}$	
Rectified Flow (RF) [4]	$X_t = (1-t)X_0 + tZ$ $t \in [0,1]$	

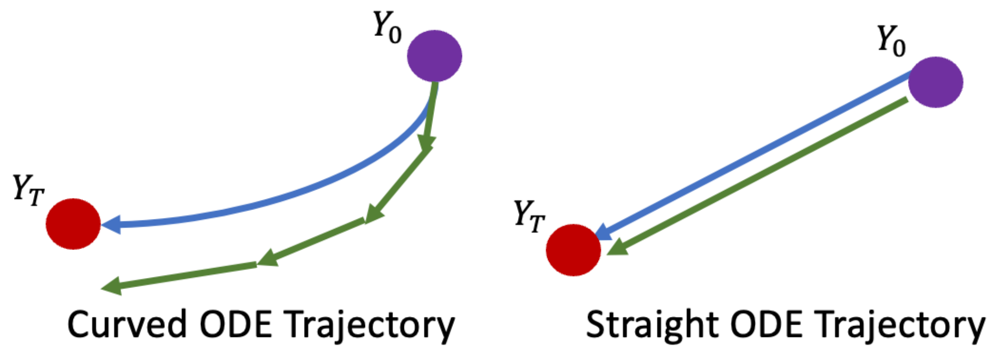


Figure from [1]

[1] LZMPL, InstaFlow: One Step is Enough for High-Quality Diffusion-Based Text-to-Image Generation, ICLR 2024.

1-Step Mapping Function from Multi-step

- For PFODE reverse process of multi-step diffusion models

$$\mathrm{d}Y_t = v(Y_t, t)\mathrm{d}t, Y_0 \sim q_T$$

  the corresponding 1-step mapping function (by integral) is

$$f(Y_t, t) = Y_{T-\delta} = X_\delta \approx X_0, \forall t \in [0, T - \delta]$$

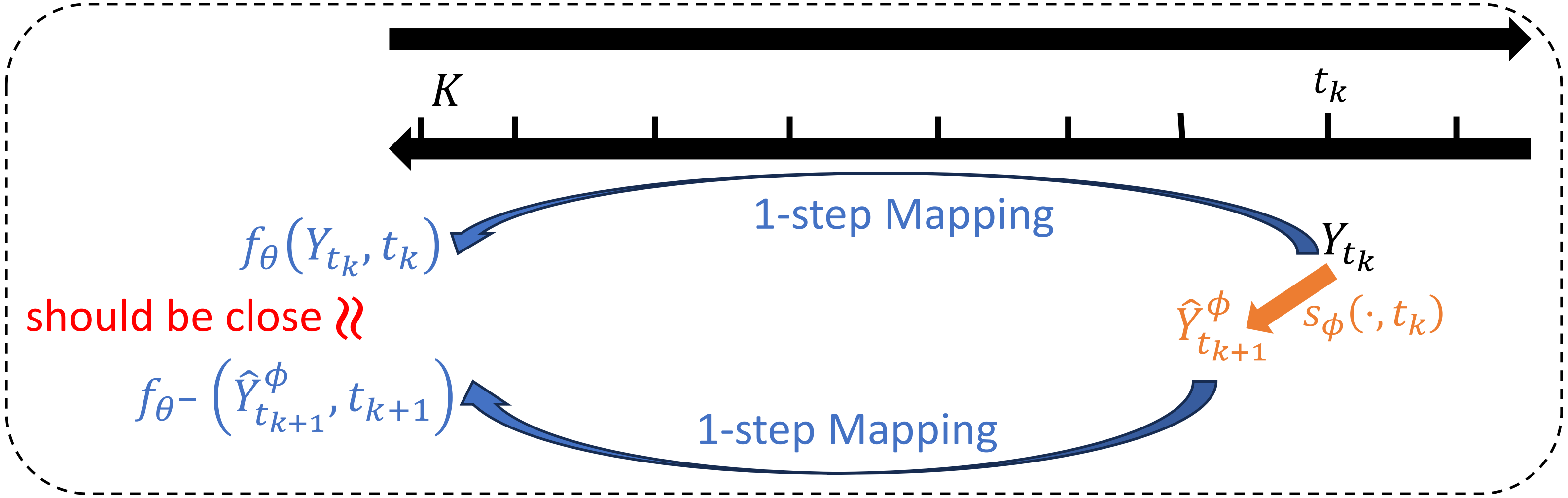
to avoid Score blow-up

- Use NN $f_\theta(Y_t, t)$ to approximate 1-step mapping function f

What is a Good Optimization Objective?

- Consistency distillation to learn good 1-step mapping [1]

$$\mathcal{L}_{\mathrm{CD}}^{K}(\boldsymbol{\theta}, \boldsymbol{\theta}^{-}; \boldsymbol{\phi}) := \mathbb{E}_{X_0}\left[\mathbb{E}_{Y_{t_k}|X_0}\left\| \boldsymbol{f}_{\boldsymbol{\theta}}(Y_{t_k}, t_k) - \boldsymbol{f}_{\boldsymbol{\theta}^{-}}\left(\hat{Y}_{t_{k+1}}^{\phi}, t_{k+1}\right)\right\|_2^2\right]$$



- Minimize K s.t. $W_2^2(f_\theta(\mathcal{N}(0, \sigma_T^2 I_d), 0; K), q_0) \leq \epsilon_{W_2}^2$

[1] SDCS, Consistency Models, ICML 2023.

[1] LCF, Sampling is as easy as keeping the consistency: convergence guarantee for consistency models , ICML 2024
[2] DCWY, Theory of consistency diffusion models: Distribution estimation meets fast sampling, ICML 2024
[3] LHW, Towards a mathematical theory for consistency training in diffusion models, AISTATS 2025
[4] YJV**L**, Improved Discretization Complexity Analysis of Consistency Models: Variance Exploding Forward Process and Decay Discretization Scheme, ICML 2025
[5] YZJC**L**, Elucidating Rectified Flow with Deterministic Sampler: Polynomial Discretization Complexity for Multi and One-step Models, Arxiv.

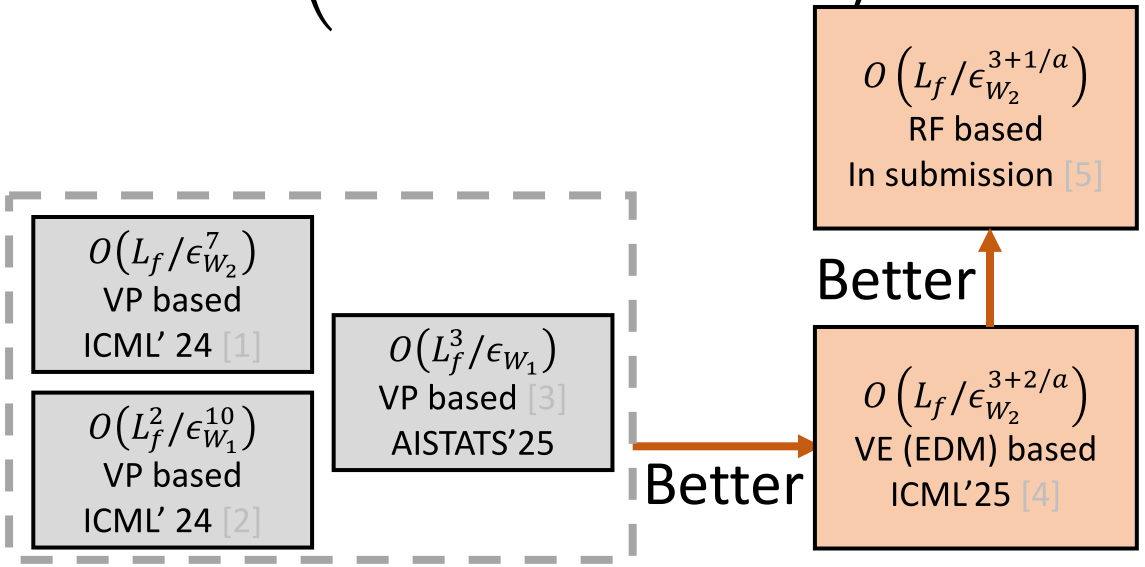
Similar Balance

- Theorem. For 1-step generation models,

$$W_2^2\left(f_\theta(\mathcal{N}(0,\sigma_T^2 I_d),T-\delta)\,,q_0\right) \le \frac{m_T}{\sigma_T^2} + \frac{L_f^2 (T/\delta)^{\frac{2}{a}}}{K^2\delta^4} + \sigma_\delta^2 \le \epsilon_{W_2}^2$$

- Then it requires discretization complexity $K = O\left(L_f (T/\delta)^{\frac{1}{a}} / (\delta^2 \epsilon_{W_2})\right)$

	T: $\frac{m_T}{\sigma_T^2} \le \epsilon_{W_2}^2$	δ: $\sigma_\delta^2 \le \epsilon_{W_2}^2$	K: $O(L_f (T/\delta)^{\frac{1}{a}} / (\delta^2 \epsilon_{W_2}))$
VP	$\log(1/\epsilon_{W_2})$ √	$\epsilon_{W_2}^2$ ×	$O\left(L_f/\epsilon_{W_2}^{5+2/a}\right)$
VE (EDM)	$1/\epsilon_{W_2}$ ×	ϵ_{W_2} √	$O\left(L_f/\epsilon_{W_2}^{3+2/a}\right)$
RF	1 √	ϵ_{W_2} √	$O\left(L_f/\epsilon_{W_2}^{3+1/a}\right)$



Conclusions

- Pretraining: Efficient Multi-manifold MoG Model
  - Empirical: Much less parameters with good enough performance
  - Theoretical: Estimation error escape the curse of dimensionality
- Fine-tuning: Good Sharing Latent Guarantees Few-shot Efficiency
  - Model the sharing scheme between pretraining and few-shot fine-tuning
  - Show effect of latent quality on estimation and optimization
- Sampling: Complexity for Multi-step Diffusion Models
  - Unified framework for sampling complexities of VP, VE, RF models
- Discretization: Complexity of 1-step Models in Training Phase
  - Support good performances of RF models

Future Work

- Pretraining Phase
  - SOTA Results with Multi-manifold MoG Modeling and Fewer Parameters
  - Global Optimization Guarantee and Generalization Mechanism
- Few-shot Fine-tuning Phase
  - Multi-task Meta-learning and Few-shot Fine-tuning Framework and Analysis
- Sampling Process of Multi-step Diffusion Models
  - Conditional Generation: Analysis of influence additional guidance
- Learning Process of 1-Step Generative Models
  - With the simplified MoG latent of Multi Subspace MoG modeling, better training and SOTA Results

Thanks!

Shuai Li

- Associate Professor
- Shanghai Jiao Tong University
- Research: RL/ML theory
- https://shuaili8.github.io/

Students:

Ruofeng Yang

Zhijie Wang

Yongcan Li

Zhaoyu Zhu

Collaborators:

Bo Jiang

Baoxiang Wang

Cheng Chen

Ruinan Jin

Questions?