上海交通大学
约翰·霍普克罗夫特
计算机科学中心
John Hopcroft Center for Computer Science

上海创智学院
SHANGHAI INNOVATION INSTITUTE
2024

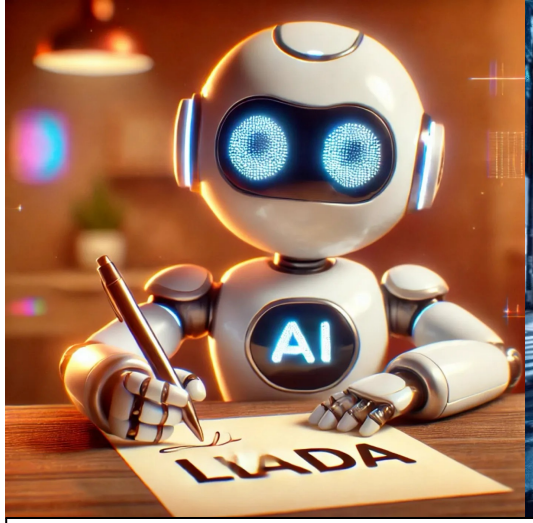
Learning Process & Sampling Complexity of Diffusion Models

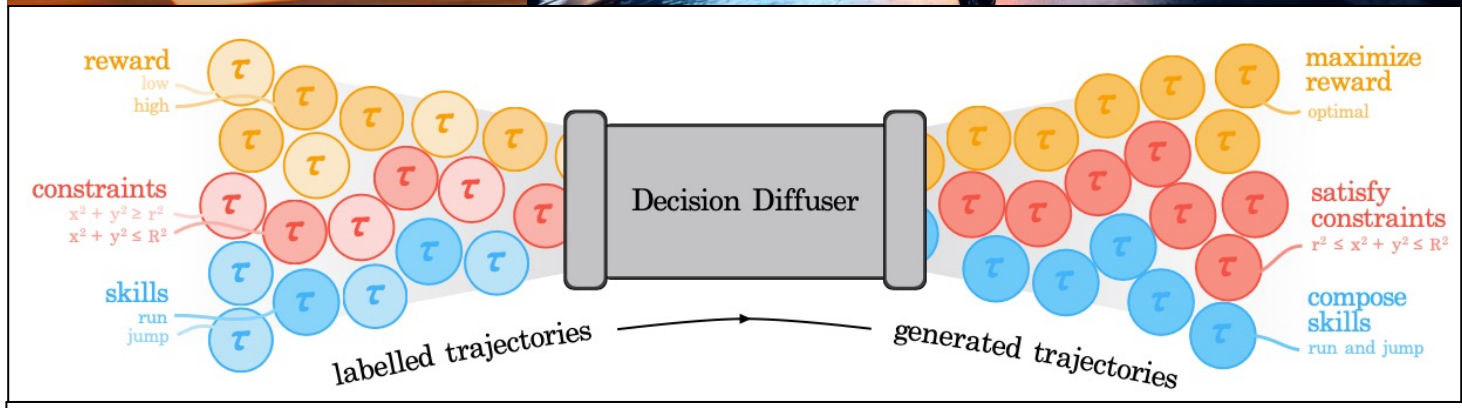
Shuai Li

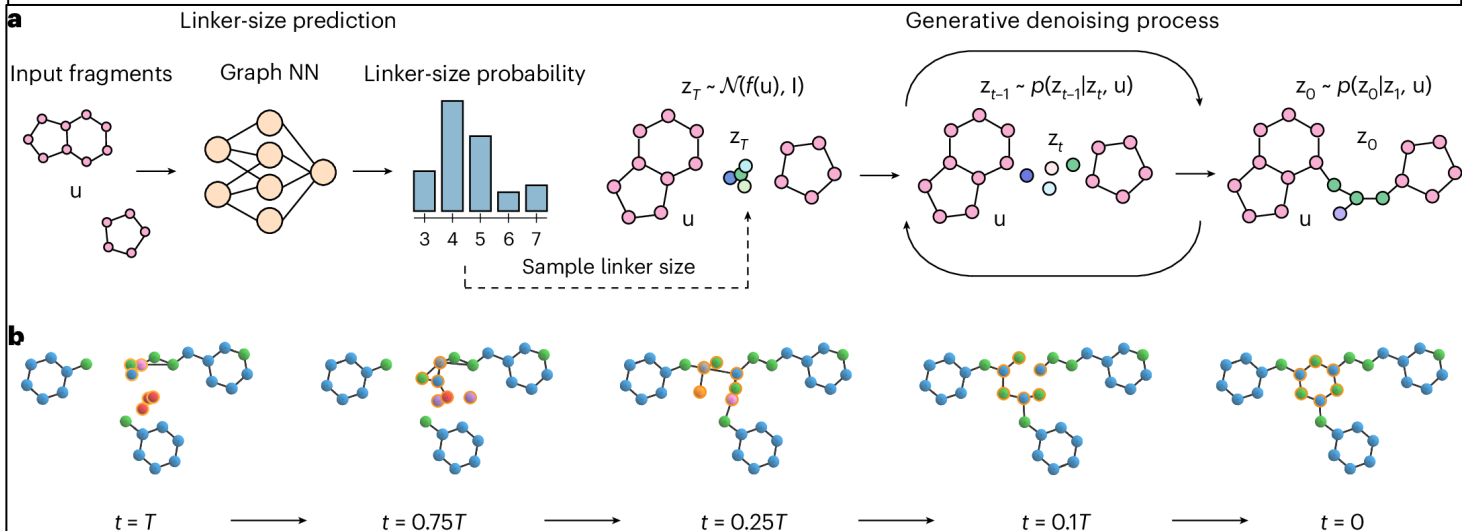
2026.07

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, NeurIPS 2025 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

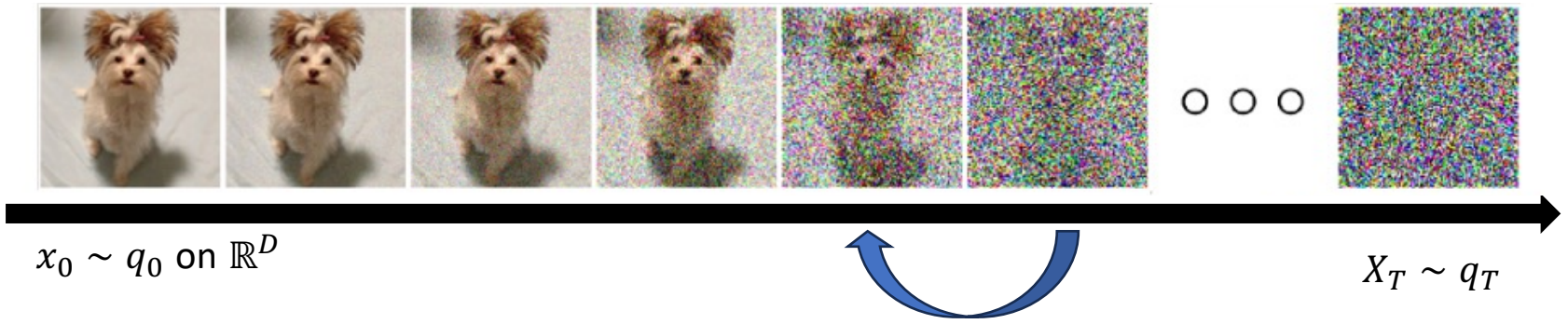
- Solid theoretical foundation helps efficient training & fast sampling:

- 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.

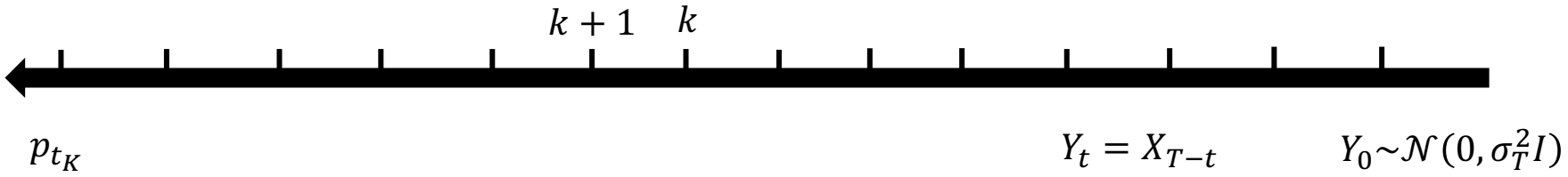
Paradigm of Multi-step Diffusion Models

Forward Process



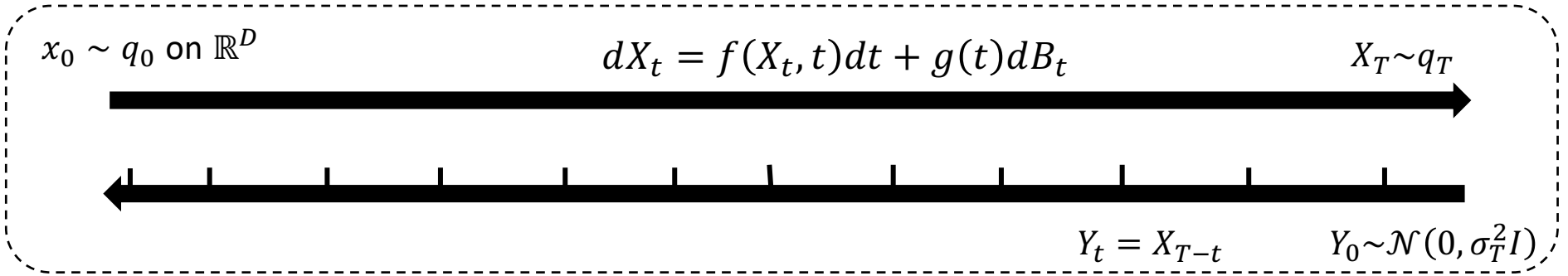
Core Problem 1: Training Process to Learn Denoising

Reverse Process



Core Problem 2: Sampling Complexity K

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:

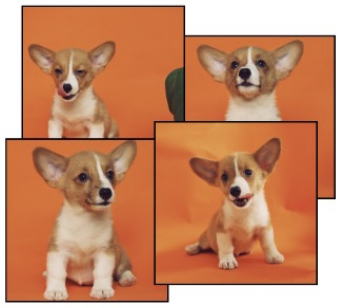
conditional distribution known

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

Overview

- **General information-sharing paradigm for few-shot finetuning**
- Mean flow with efficient data modeling
- Understanding of the guidance strength in the sampling process

Few-shot Fine-tuning



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

e.g. 882M

- Estimation error $O(n_\mathrm{s}^{-\frac{2}{d}})$ — Tolerable!

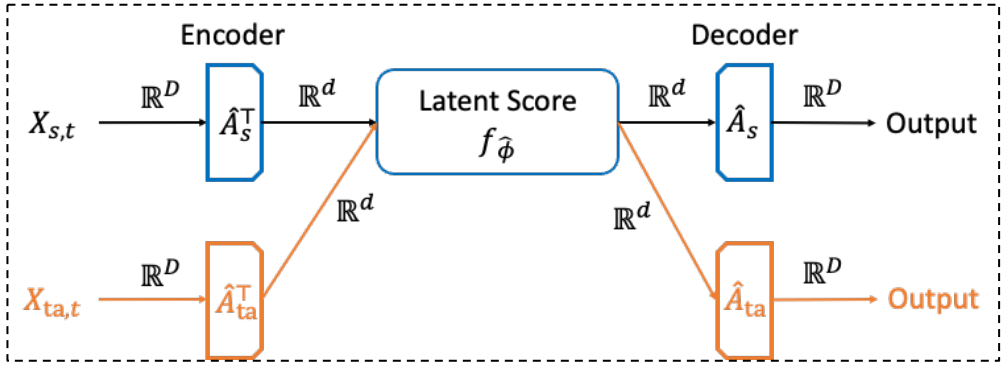
- Fine-tune with limited target data (~10 images): $\{X_{\mathrm{ta},i}\}_{i=1}^{n_\mathrm{ta}} \sim q_0^{\mathrm{ta}}$
- $\min_{s \in \text{Target } \mathfrak{P}} \hat{\mathcal{L}}_\mathrm{ta}(s) = \frac{1}{n_\mathrm{ta}} \sum_{i=1}^{n_\mathrm{ta}} \frac{1}{T-\delta} \int_\delta^T \mathbb{E}_{X_t|X_0=X_{\mathrm{ta},i}}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

e.g. 1.5M 0.17%

- Estimation error $O(n_\mathrm{ta}^{-\frac{2}{d}})$ — Meaningless!

Previous: Information-sharing Model Design

- Empirical works share most parameters and fine-tune key parameters
- Assumption. Data admit linear structure $X_s = A_s z, X_{ta} = A_{ta} z, z \in \mathbb{R}^d$ and share latent space/score q_t^{Latent}



- Then the score function is

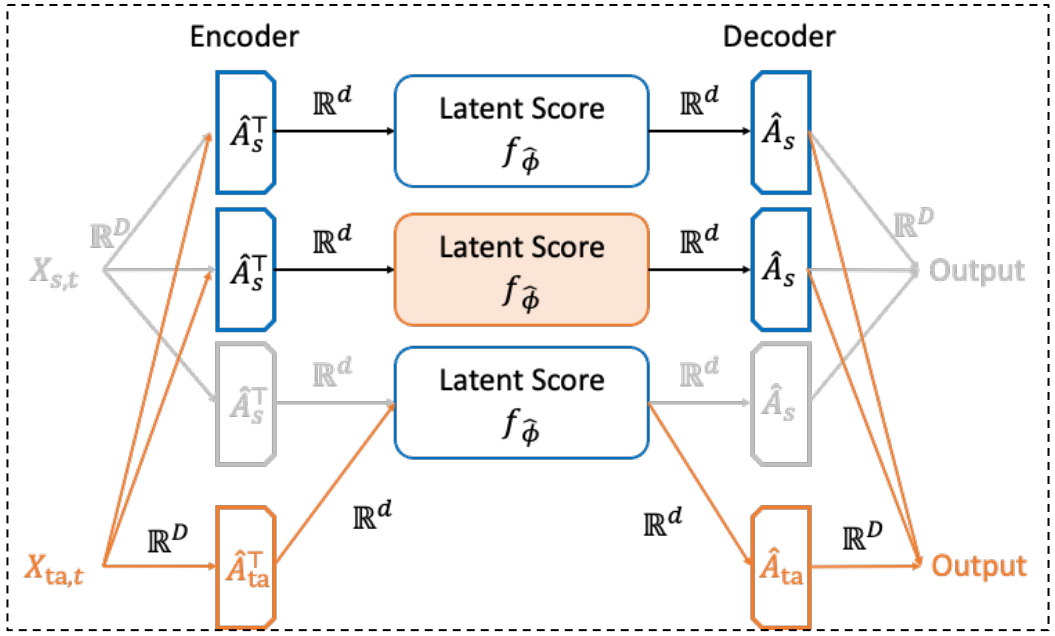
$$\nabla \log q_t^{\text{ta}}(X) = A_{\text{ta}} \nabla \log q_t^{\text{Latent}}\left(A_{\text{ta}}^\top X\right) - \frac{1}{\sigma_t^2}\left(I_D - A_{\text{ta}} A_{\text{ta}}^\top\right) X$$

Shared Latent Score

- OLLELSZ, Few-Shot Image Generation via Cross-Domain Correspondence, CVPR 2021.
- YHCJWL, Few-shot diffusion models escape the curse of dimensionality, NeurIPS 2024.
- YLJCL, Evaluating the Role of Great Pre-trained Diffusion Models in Few-shot Phase: Warm-up and Acceleration, UAI 2026.

More General Information-sharing Paradigm

- Concepts are composable [1], localized [2], and partially shared [3].
  - styles, colors, structure, …
- Assumption. For every latent subspace and its latent score function, any of the following satisfies:
  - Target share both encoder and latent
  - Target share encoder but no latent
  - Target share latent but no encoder



[1] PYAW, Orthogonal Adaptation for Modular Customization of Diffusion Models, CVPR 2024.
[2] LFZZZLZZC, Cones: Concept Neurons in Diffusion Models for Customized Generation, ICML 2023.
[3] KZZSZ, Multi-Concept Customization of Text-to-Image Diffusion, CVPR 2023.
[4] WGBHKW, MineGAN: Effective Knowledge Transfer From GANs to Target Domains With Few Images, CVPR 2020.
[5] KZZSZ, Multi-Concept Customization of Text-to-Image Diffusion, CVPR 2023.
[6] LYL, Your Few-shot Diffusion Models Secretly Localize to a Small Subspace: An Explanatory Framework in the Latent Space, in submission.

Theoretical Efficiency Guarantee

- Theorem. The estimation error for few-shot fine-tuning is

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

$$\leq O\left(\left(\sqrt{\mathrm{d\_z\_unshared/d\_z\_all}}+\sqrt{\mathrm{d\_A\_unshared/d\_A\_all}}\right)n_{\mathrm{ta}}^{-\frac{1}{2}}+n_s^{-\frac{2}{d}}\right)$$

  - Assume every subspace has GMM structure

- Escape curse of dimensionality
- Can recover previous result

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

- LY**L**, Your Few-shot Diffusion Models Secretly Localize to a Small Subspace: An Explanatory Framework in the Latent Space, in submission.

Finetune Less Params w/ Good Performance

Method	Finetune Params	CLIP-T ↑
Textual Inversion [1]	768	0.2258±0.0238
DreamBooth [2] (with LoRA, r = 1)	99.6K	0.2752 ±0.0125
Ours [3] (only recalibrate unshared ones*)	2.13K 2.14%	0.2743 ±0.0088 (Comparable!)

Only the structure was adjusted

* We normalize the frozen latent features and identify a subspace in which the source and target distributions are best separated (e.g., using LDA), treating its complement as shared.

[1] GAAPBCC, An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, ICLR 2023.
[2] RLJPRA, DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, CVPR 2023.
[3] LY**L**, Your Few-shot Diffusion Models Secretly Localize to a Small Subspace: An Explanatory Framework in the Latent Space, in submission.
[4] CCCXYTLLH, Deep Compression Autoencoder for Efficient High-resolution Diffusion Models, ICLR 2025.

Overview

- General information-sharing paradigm for few-shot finetuning
- **Mean flow with efficient data modeling**
- Understanding of the guidance strength in the sampling process

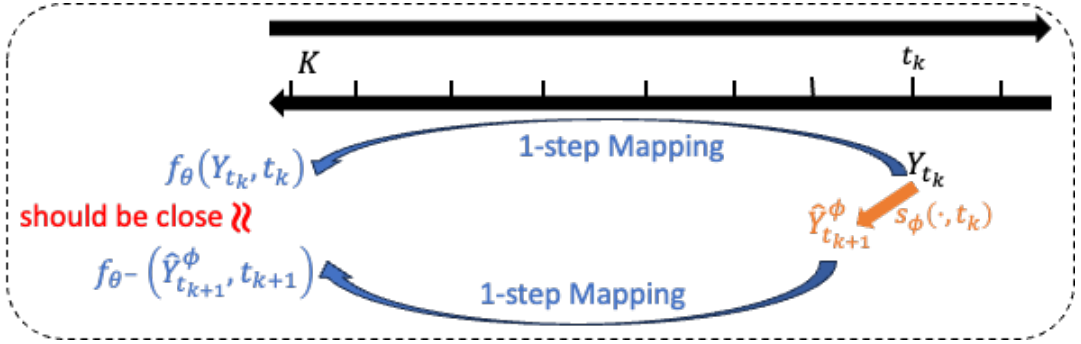
From Consistency Model to Mean Flow

- Consistency model learns the 1-step mapping function

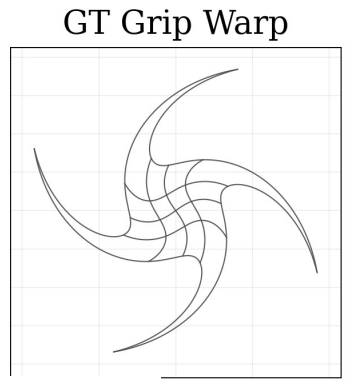
$$f(Y_t, t) = Y_{T-\delta} = X_\delta \approx X_0, \forall t \in [0, T-\delta]$$

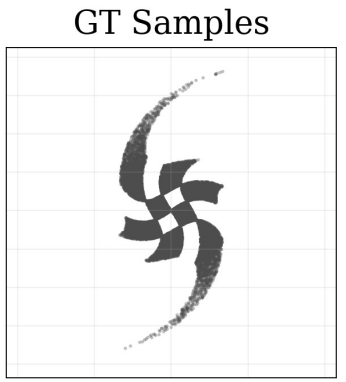
from PFODE reverse process of multi-step diffusion models

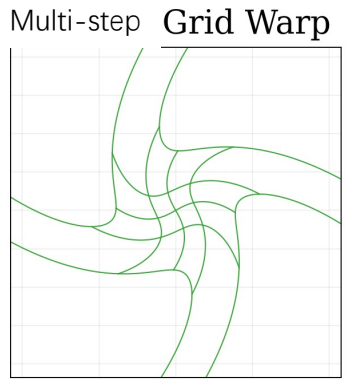
$$\mathrm{d}Y_t = v(Y_t, t)\mathrm{d}t, Y_0 \sim q_T$$



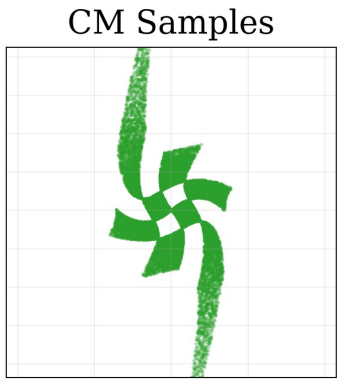
- Its learning of the ODE integral is based on the given multi-step diffusion model

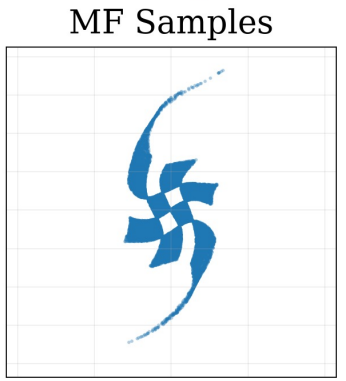






discretization leads to large error





Mean Flow Model

- Mean flow model directly learns the integration on arbitrary intervals, thus ⟺ learn the average velocity on arbitrary intervals
- Average velocity

$$u(Y_t, r, t) \coloneqq \frac{1}{t-r}\int_r^t v(Y_\tau, \tau)d\tau$$

$$= v(Y_t, t) - (t-r)\frac{\mathrm{d}}{\mathrm{dt}}u(Y_t, r, t) = v(Y_t, t) - (t-r)(v(Y_t, t)\partial_Y u + \partial_t u)$$

$$\frac{\mathrm{d}}{\mathrm{dt}}u(Y_t, r, t) = \frac{\mathrm{dY_t}}{\mathrm{dt}}\partial_Y u + \frac{dr}{dt}\partial_r u + \frac{dt}{dt}\partial_t u, \frac{dY_t}{dt} = v, \frac{dr}{dt} = 0$$

better form w.o. integration

- The learning objective of the mean flow model is

$$\mathcal{L}_{\mathrm{MF}}(\theta) = \mathbb{E}\|u_\theta(Y_t, r, t) - u_{\mathrm{ta}}\|_2^2$$

- GDBKH, Mean Flows for One-step Generative Modeling, NeurIPS 2025.

MF has a Better Learning Objective than CM

- Training objective of MF directly bounds the generation performance

$$W_2^2(u_\theta(\mathcal{N}(0,\sigma_T^2 I_d),0,T)\,,q_0) \leq \mathcal{L}_{\mathrm{MF}}(\theta)$$

- $\mathcal{L}_{\mathrm{MF}}(\theta^*) = 0$ for some θ^* in a moderately good function class

- Training objective of CM has a multi-step prior bias, thus leads to

$$W_2^2(f_\theta(\mathcal{N}(0,\sigma_T^2 I_d),0)\,,q_0) \leq \mathcal{L}_{\mathrm{CD}}^K(\theta,\theta^-;\phi) \text{ + prior bias}$$

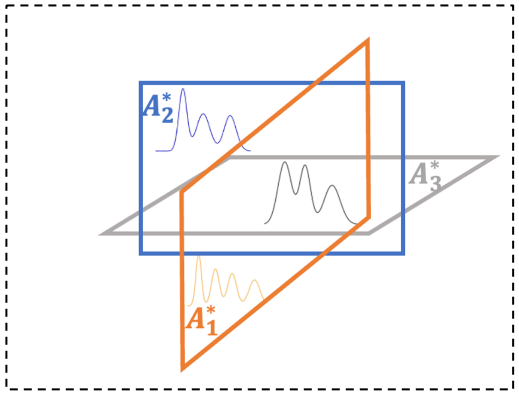
- CHLY**L**, Mean Flows Learn Intrinsic Flow Maps: Dimension-Efficient Guarantees under Low-Rank Gaussian Mixtures, In submission.

Data Modeling: Intrinsic Gaussian Mixture Flow (IGMF)

- One-step generation is very difficult to train [1]
- Data modeling helps in efficient training, especially in one-step generation [2][3][4]

- Inspired by efficient GMM modeling for multi-step generation [5]

$$X \sim \sum_{\ell=1}^{L} \pi_\ell \sum_{m=1}^{M} \pi_{\ell,m} \mathcal{N}\left(\cdot; A_\ell \mu_{\ell,m}, A_\ell \Sigma_{\ell,m} A_\ell^\top\right)$$



[1] FHLA, One Step Diffusion via Shortcut Models, ICLR 2025 Oral.
[2] LSYWYL, SubFlow: Sub-mode Conditioned Flow Matching for Diverse One-Step Generation, Arxiv.
[3] KCY, Align Your Trajectory Tangent: Training Better Consistency Models via Manifold-Aligned Tangents, ICML 2026.
[4] HLWME, MeanFlow Transformers with Representation Autoencoders, CVPR 2026.
[5] YLJC**L**, Multi-Subspace Multi-Modal Modeling for Diffusion Models: Estimation, Convergence and Mixture of Experts, ICLR 2026.

Data Modeling: Intrinsic Gaussian Mixture Flow (IGMF) 2

- Note the average velocity for the single Gaussian modeling has a closed analytical form

$$A_t(\lambda) = \frac{t-(1-t)\lambda}{(1-t)^2\lambda + t^2}, \quad B_t(\lambda) = \frac{-t}{(1-t)^2\lambda + t^2}, \quad g_{k,t}(\xi_k) = \left(A_t(\Lambda_k) - \frac{1}{t} I_{d_k}\right)\xi_k + B_t(\Lambda_k) b_k$$

where $A_t(\Lambda_k)$ and $B_t(\Lambda_k)$ are defined entrywise on the diagonal matrix Λ_k. Define

$$D_t(\lambda) = (1-t)^2\lambda + t^2, \quad F_{r,t}(\lambda) = \frac{r}{t} - \sqrt{\frac{D_r(\lambda)}{D_t(\lambda)}}, \quad G_{r,t}(\lambda) = (1-t)\sqrt{\frac{D_r(\lambda)}{D_t(\lambda)}} - (1-r)$$

The structures of the Gaussian velocity field and flow map are given by the following lemma.

Lemma 5.1. *If $x_0 \sim \mathcal{N}(\mu_k, \Sigma_k)$, then the velocity field $v(z_t, t)$ and flow map $X(z_t, r, t)$ of the PF-ODE satisfy*

$$v(z,t) = \frac{1}{t} z + U_k g_{k,t}(\xi_k), \quad X(z_t, r, t) = \frac{r}{t} z_t - U_k \left(F_{r,t}(\Lambda_k)\xi_{k,t} + G_{r,t}(\Lambda_k) b_k\right)$$

- CHLY**L**, Mean Flows Learn Intrinsic Flow Maps: Dimension-Efficient Guarantees under Low-Rank Gaussian Mixtures, In submission.

Data Modeling: Intrinsic Gaussian Mixture Flow (IGMF) 3

- But for the mixture of Gaussians, the flow map

$$\Phi(Y_t, r, t) \coloneqq Y_t - (t-r)u(Y_t, r, t) = \frac{r}{t}Y_t - r\sum_{\ell,m} A_{\ell,m}\int_r^t \frac{1}{s}\tilde{\gamma}_{\ell,m}(\xi_s, s)g_{\ell,m,s}(\xi_{\ell,m,s})\mathrm{d}s$$

  does not have a closed form since $\tilde{\gamma}_{\ell,m}(\xi_s, s)$ is nonlinear

- Assumption. The flow map satisfies

$$\Phi(Y_t, r, t) = \frac{r}{t}Y_t - r\sum_{\ell,m} A_{\ell,m}\int_r^t \frac{1}{s}\tilde{\gamma}_{\ell,m}(\xi_t, t)g_{\ell,m,s}(\xi_{\ell,m,s})\mathrm{d}s$$

$$= \frac{r}{t}Y_t - \sum_{\ell,m} A_{\ell,m}\,\gamma_{\ell,m}(Y_t, t)\left(F_{r,t}(\Lambda_k)A_{\ell,m}^{\top}Y_t + G_{r,t}(\Lambda_k)b_k\right)$$

- CHLY**L**, Mean Flows Learn Intrinsic Flow Maps: Dimension-Efficient Guarantees under Low-Rank Gaussian Mixtures, In submission.

Estimation Error under IGMF

- Theorem. Assume continuous enough velocity and flow map, the estimation error satisfies

$$\mathbb{E}_{q_0}\left[\left\|u_\theta(\mathcal{N}(0,\sigma_T^2 I_d),0,T)-u^*(\mathcal{N}(0,\sigma_T^2 I_d),0,T)\right\|_2^2\right] < O(\sqrt{d}n^{-\frac{2}{d+2}}).$$

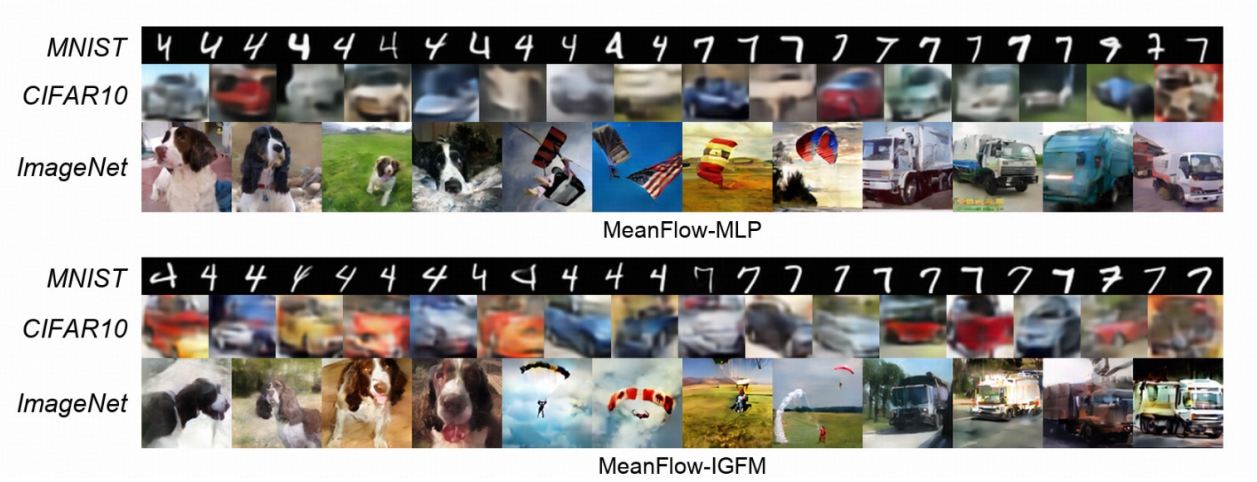
- Theorem. The estimation error under IGMF Modeling satisfies

$$\mathbb{E}_{q_0}\left[\left\|u_\theta(\mathcal{N}(0,\sigma_T^2 I_d),0,T)-u^*(\mathcal{N}(0,\sigma_T^2 I_d),0,T)\right\|_2^2\right] < O\left(d^3\sqrt{\frac{MLd}{n}}\right).$$

- CHLY**L**, Mean Flows Learn Intrinsic Flow Maps: Dimension-Efficient Guarantees under Low-Rank Gaussian Mixtures, In submission.

IGMF is a Good Modeling

Dataset	NN Model	# of Parameters	LPIPS↓	CLIP↑
MNIST	MLP	2.11M	0.2452	0.2858
	IGMF	0.028M 1.33%	0.2369 ↓3.4%	0.2886 ↑1.0%
CIFAR10	MLP	10.05M	0.5643	0.2594
	IGMF	0.106M 1.05%	0.4798 ↓15.0%	0.2634 ↑1.5%



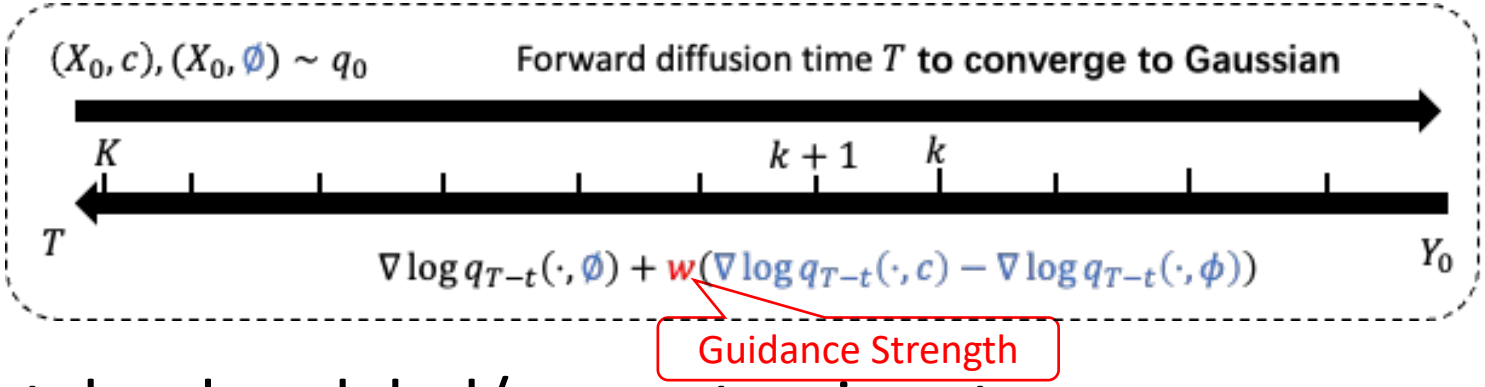
- IGMF modeling uses much less parameters and is much easier to train well
- Can realize 1-step generation than previous MoLR-MOG (20x acceleration)

- CHLY**L**, Mean Flows Learn Intrinsic Flow Maps: Dimension-Efficient Guarantees under Low-Rank Gaussian Mixtures, In submission.

Overview

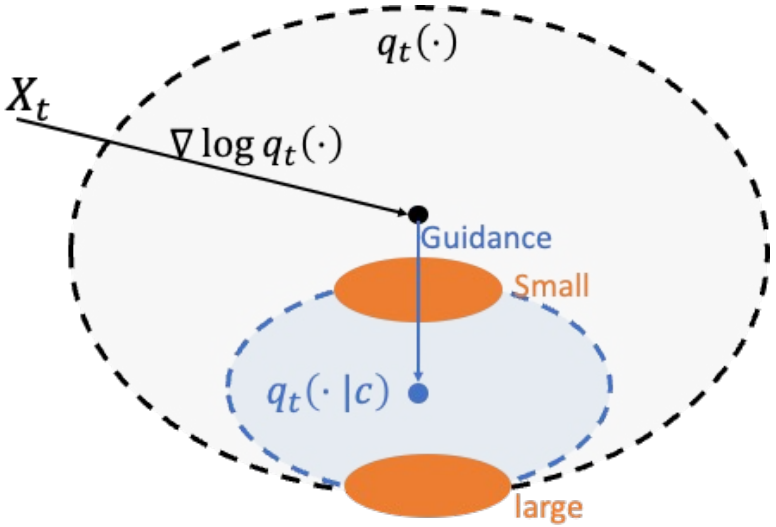
- General information-sharing paradigm for few-shot finetuning
- Mean flow with efficient data modeling
- **Understanding of the guidance strength in the sampling process**

Guidance



- Generation tasks usually take class-label/prompt as input
- Learn $\nabla \log q_t(\cdot\,|c)$ directly mainly apply to discrete/independent c
- For continuous c, $\nabla \log q_t(\cdot\,|c) = \nabla \log q_t(\cdot) + \nabla \log q_t(c|\,\cdot)$
- Use $\nabla \log q_t(\cdot, \emptyset) + w(\nabla \log q_t(\cdot, c) - \nabla \log q_t(\cdot, \emptyset))$ as guided score

Classfier-free Guidance (CFG)



- Small guidance ⟶ Bad alignment
- Large guidance ⟶ OOD & Modal Collapse

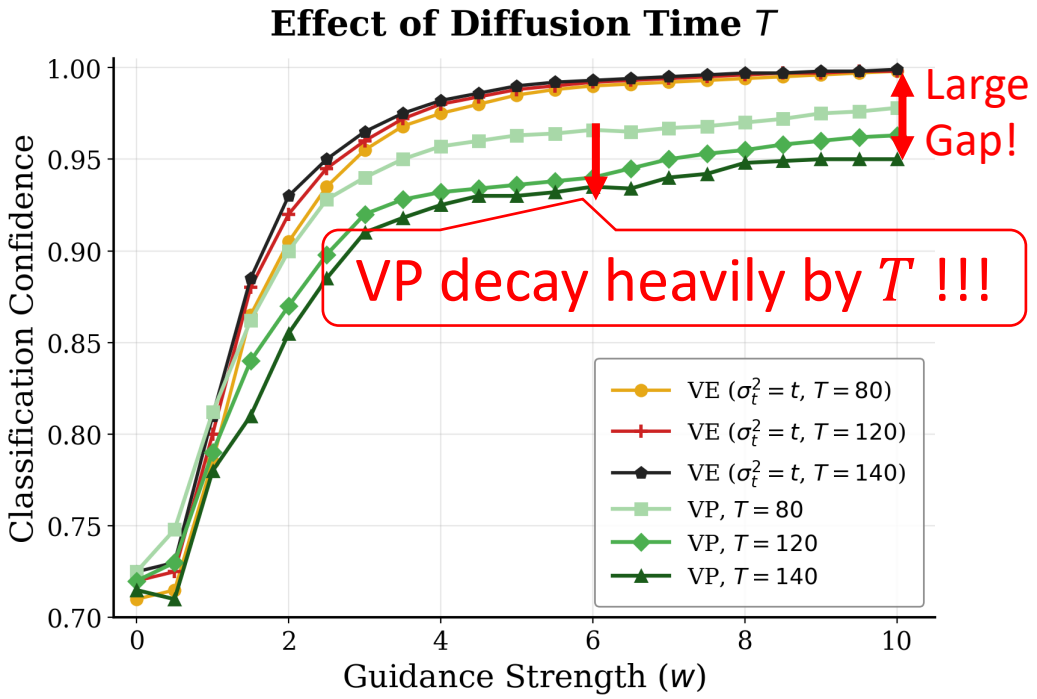
What is the influence of guidance strength?

Measurement: Classficiation Confidence

- $\mathbb{P}(c|X)$ is the classification confidence
- Ideally $\mathbb{P}(c|X) = 1$ if X belongs to class c

- Compare VP vs VE w/ 3GMM
- Assume accurate score function
- VP is much worse than VE

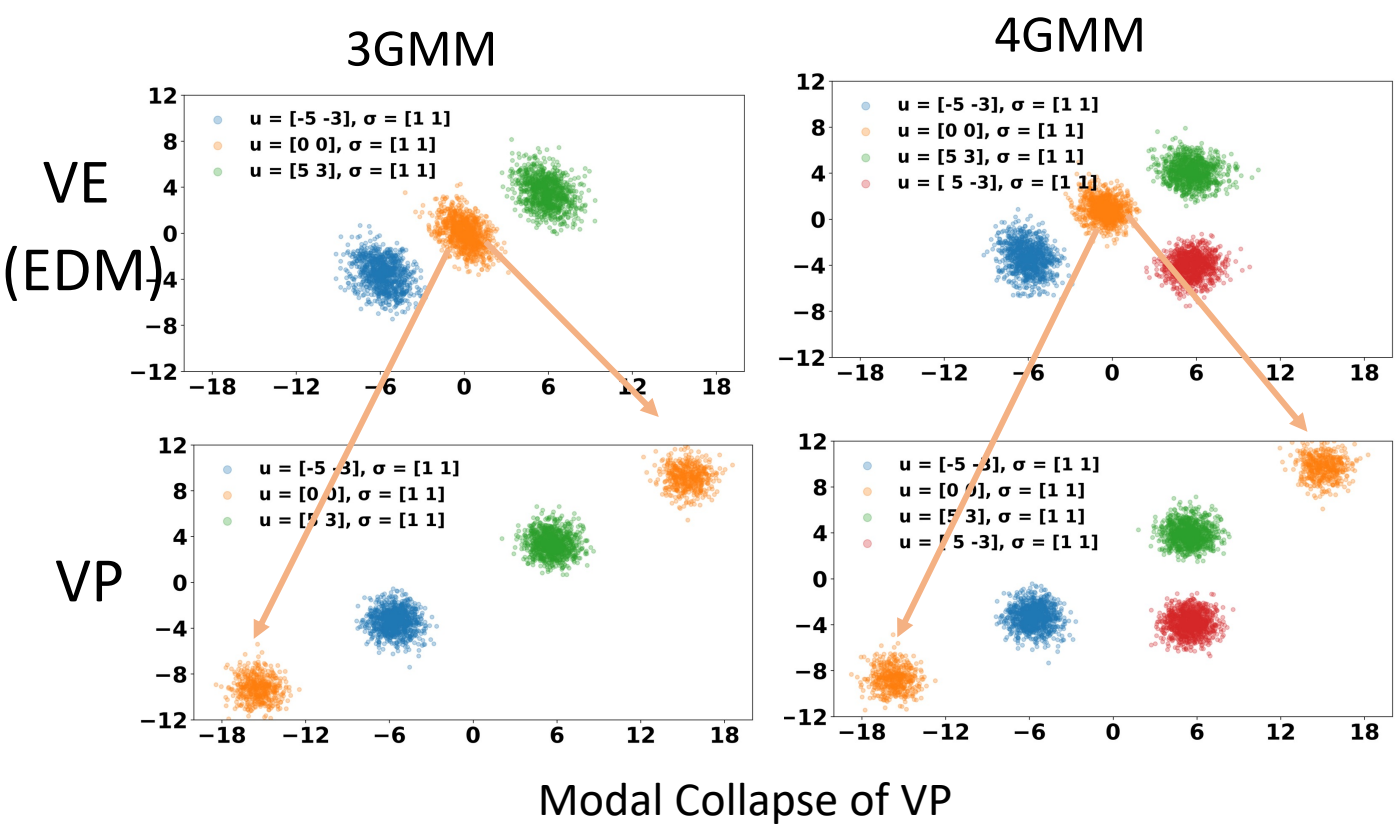
Forward Process	Classification Confidence
VP	$1 - \eta^{-e^{-T}}(\log \eta)^2$
VE-EDM	$1 - \eta^{-1}(\log \eta)^2$



[1] WCLWW, Theoretical insights for diffusion guidance: A case study for Gaussian mixture models, ICML 2024

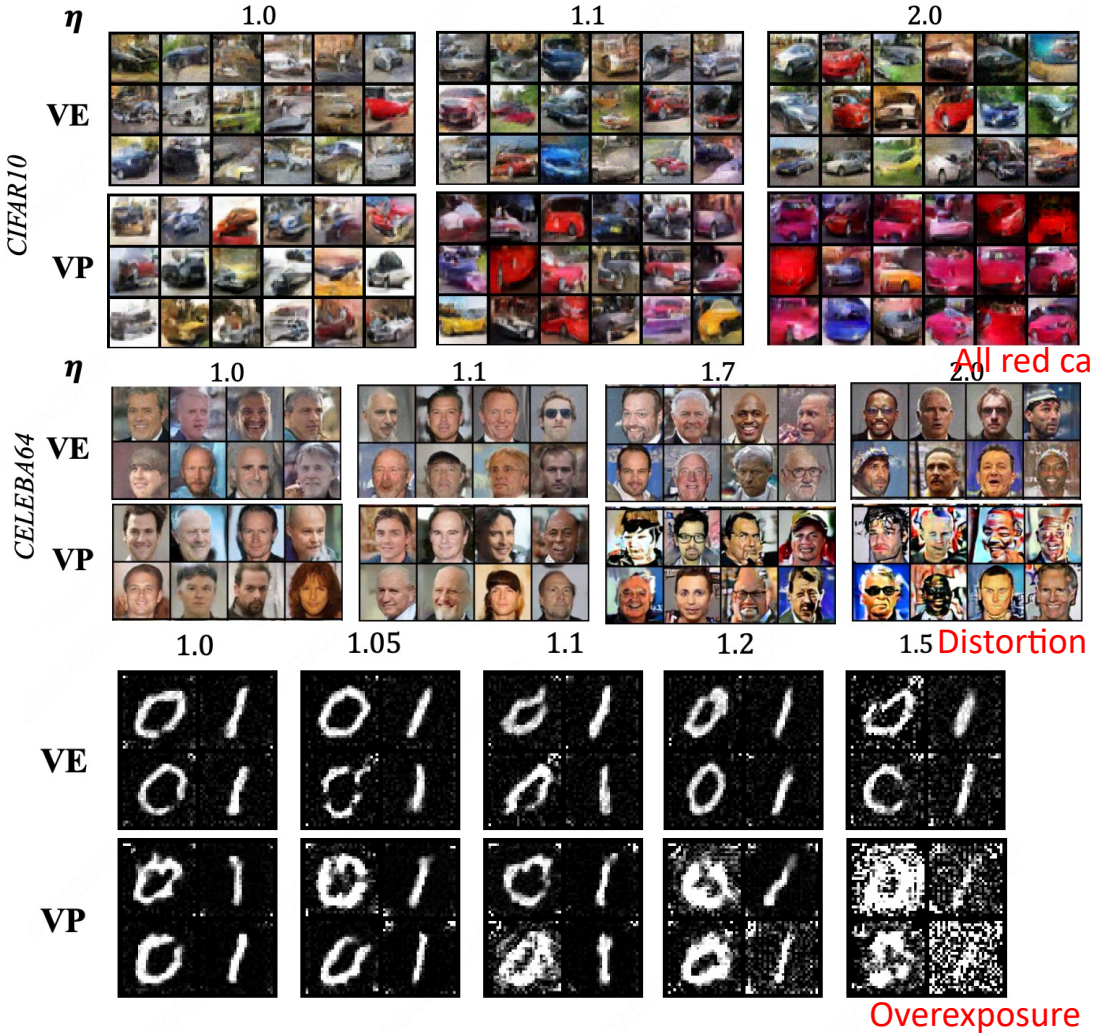
[2] YCJCL, Elucidating Guidance in Variance Exploding Diffusion Models: Fast Convergence and Better Diversity (ICLR 2026 DeLTa Workshop)

Diversity/Modal Collapse of VP w/ large guidance



Modal Collapse of VP

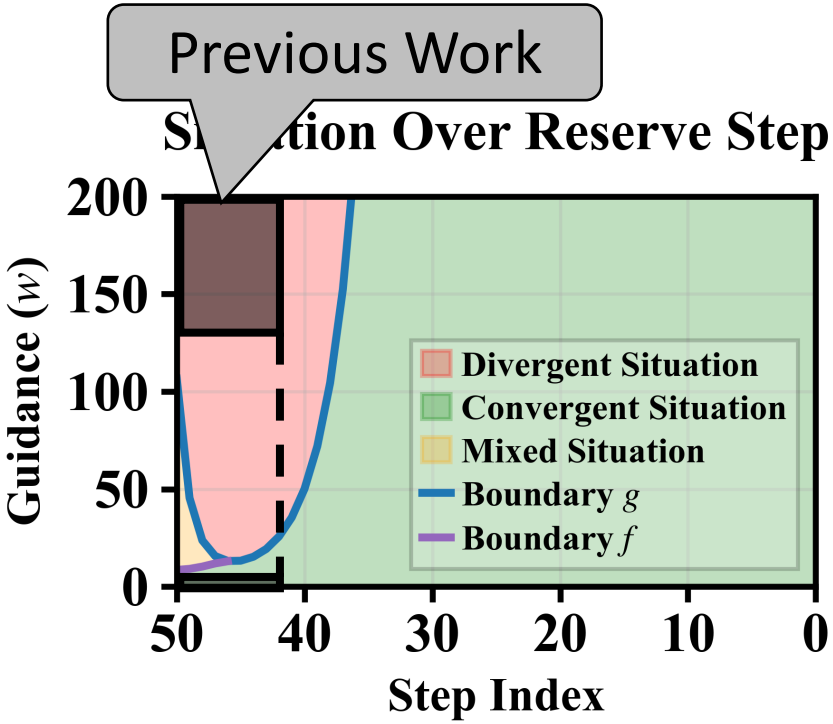
What happens?



Full Guidance Characterization of VP for GMM

- 3GMM and target central 0 mode

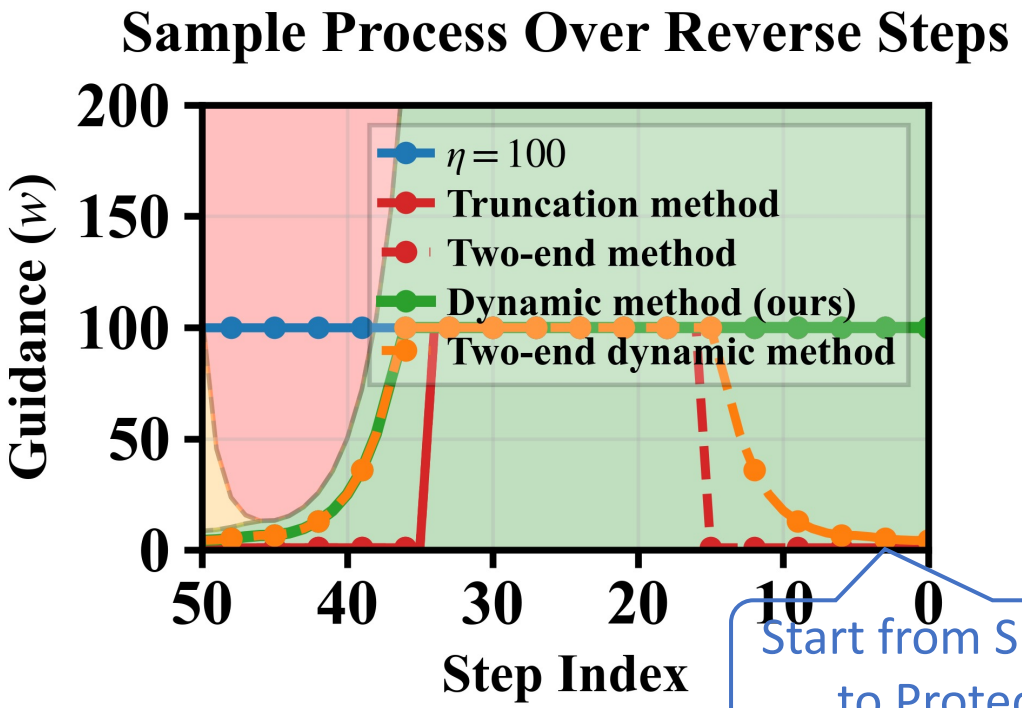
- 3-situation full characterization:
  - Convergent situation: Closer to center 0
  - Divergent situation: Farther from center 0
  - Mixed: Could be both

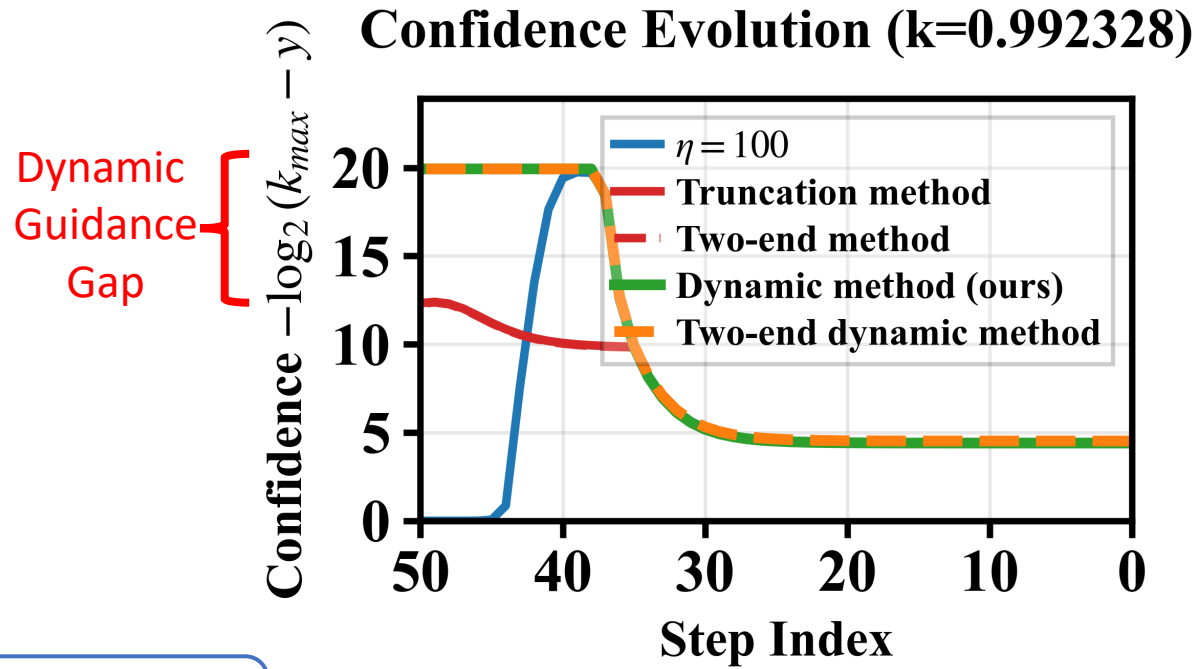


[1] WCLWW, Theoretical insights for diffusion guidance: A case study for Gaussian mixture models, ICML 2024.
[2] QYYL, On the Necessity of Guidance Decay: From Three-Phase Analysis in Gaussian Mixture Models to Dynamic Optimization (In submission).

Schedule of Guidance Strength

- Safe-region guidance leads to high classification confidence
- Dynamic/larger guidance better than truncation method

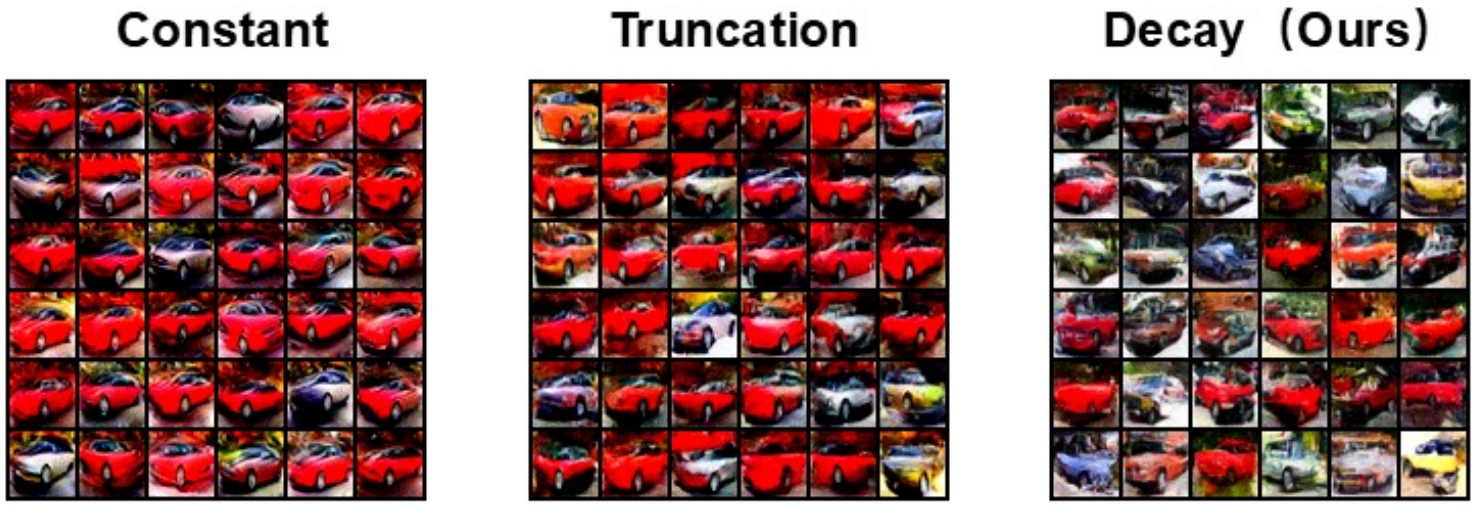




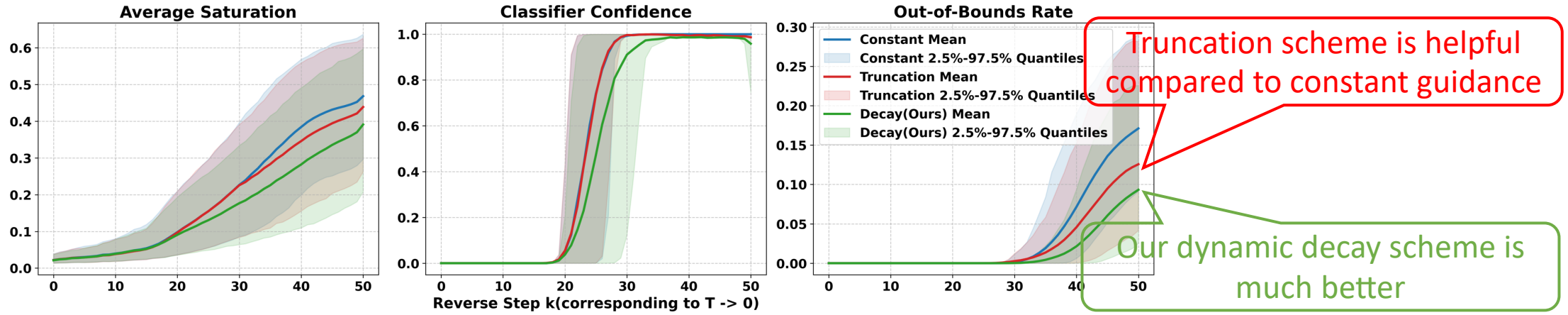
Start from Small Guidance to Protect Diversity

- QYYL, On the Necessity of Guidance Decay: From Three-Phase Analysis in Gaussian Mixture Models to Dynamic Optimization (In submission).

Real-world Experiments with Dynamic Guidance

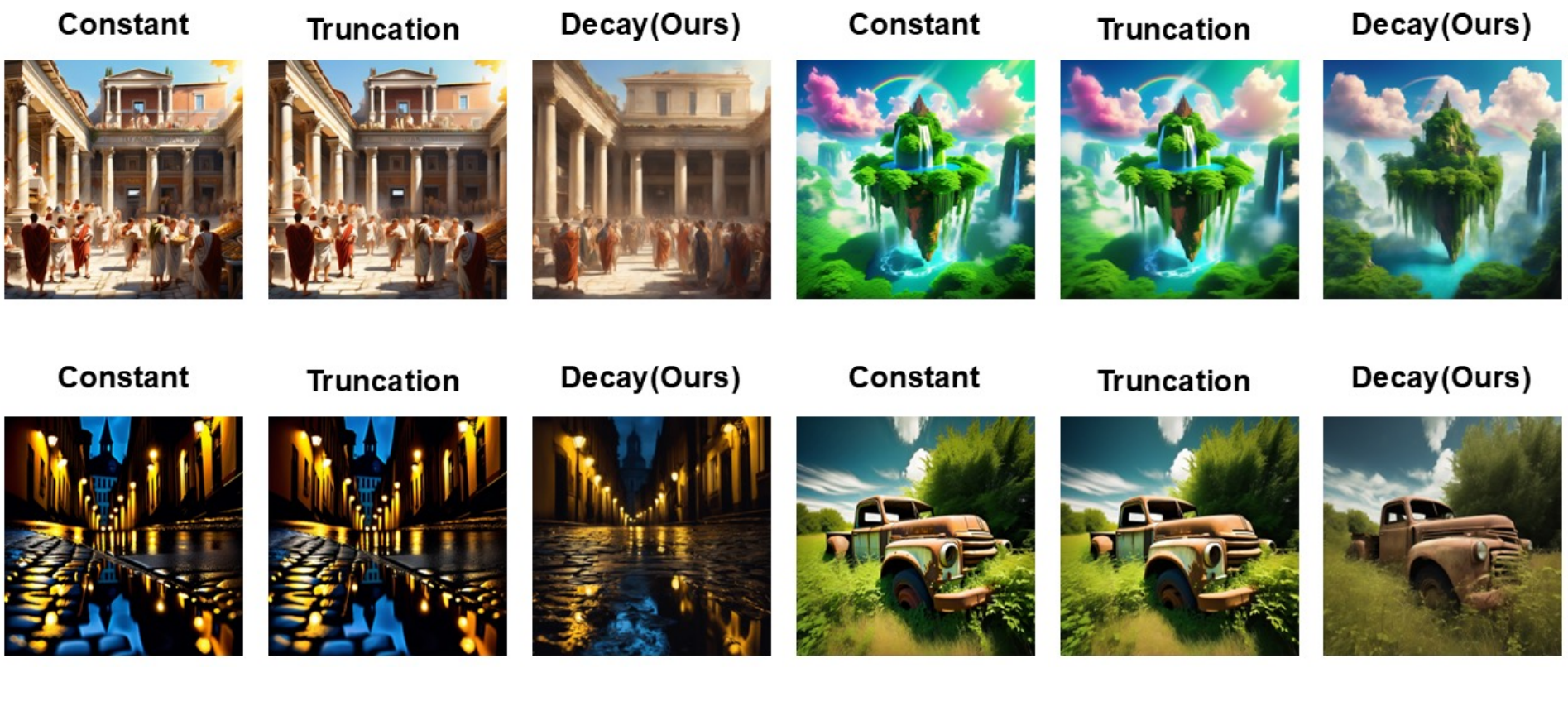


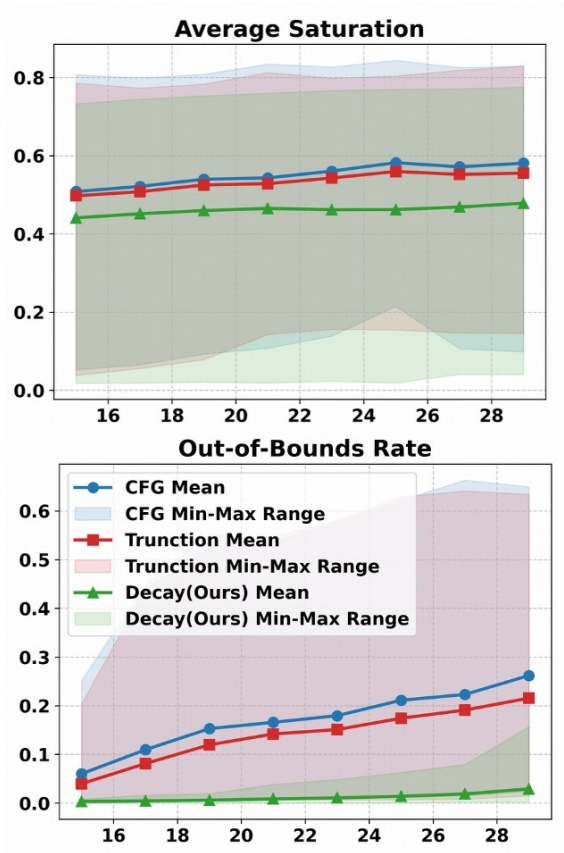
Dynamic Decay Guidance
Has Lower Saturation & More Diversity



- QYYL, On the Necessity of Guidance Decay: From Three-Phase Analysis in Gaussian Mixture Models to Dynamic Optimization (In submission).

Real-world Experiments with Dynamic Guidance





- QYY**L**, On the Necessity of Guidance Decay: From Three-Phase Analysis in Gaussian Mixture Models to Dynamic Optimization (In submission).

Conclusions

- Fine-tuning: Good Sharing Latent Guarantees Few-shot Efficiency
  - A unified view of map adaptation and latent adaptation through the same information-sharing principle
  - Frozen-VAE adaptation of only unshared factors
- MeanFlow: Intrinsic Structure Enables Efficient 1-Step Generation
  - Average-velocity regression without consistency-induced discretization error
  - With the Multi Subspace MoG modeling, better estimation
- Guidance: Dynamic CFG Balances Alignment and Diversity
  - Three-regime characterization of VP; faster confidence convergence and better diversity for VE
  - Dynamic schedules achieve high confidence with lower saturation, fewer OOD samples, and less mode collapse

Future Work – Continuous Diffusion

- Few-Shot Fine-tuning Phase
  - Better shared–unshared factor discovery
  - Multi-task transfer with guarantees under latent and model mismatch
- MeanFlow
  - Richer intrinsic data modeling beyond low-rank Gaussian mixtures
  - Scalable lightweight 1-step models with optimization and generalization guarantees
- Guidance
  - Extend CFG analysis beyond VP/VE and Gaussian mixture models
  - Based on current latent and prompt, achieve adaptive guidance schedules balancing alignment, diversity, and OOD risk

Future Work – Discrete Diffusion/dLLM

- Pretraining Phase
  - Pretraining with Text Data Structure (such as Chain MRF, Tree Structure)
  - Estimation Error with Data Structure to Escape the Curse of Dimensionality
  - Analysis for Fundamental Difference between continuous and discrete dLLM
- dLLM Agent: Parallel Decoding Mechanism Leads to Higher Efficient but Less Stable Tool Call
  - AR Distill SFT
  - Tree Based Step-by-Step GPRO Post Training Algorithm

Thanks!

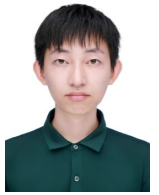
Students:


Ruofeng Yang


Yongcan Li


Xinyu Che


Yiyu Qiu


Junhao He

Collaborators:


Bo Jiang


Cheng Chen

Shuai Li

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

Questions?