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

Diffusion Models Meet Representation Learning and Manifold Learning

Ruofeng Yang

Paradigm of Multi-step Diffusion Models

Forward Process

$x_0 \sim q_0$ on $\mathbb{R}^d$

$X_T \sim q_T$

Reverse Process

$k+1$ k

p_{t_K}

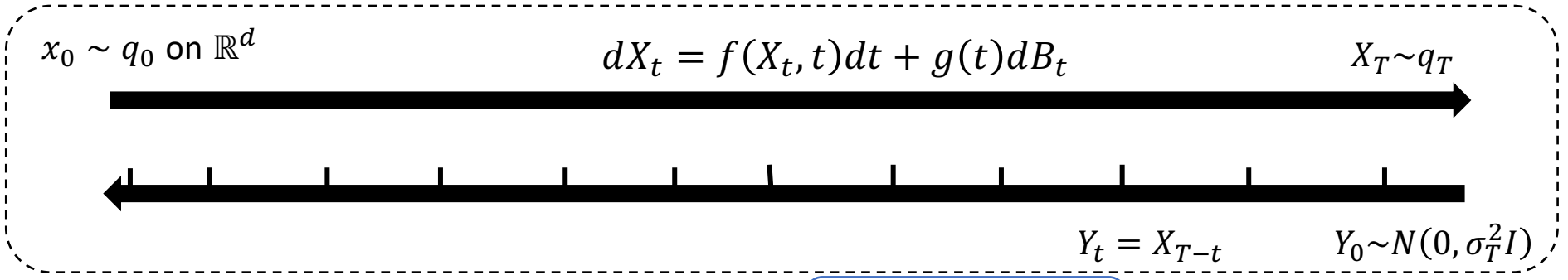
$Y_t = X_{T-t}$

$Y_0 \sim N(0, \sigma_T^2 I)$

Overview

- The Relationship between Diffusion Models and SSL
- How to use Data Structure to improve Performance
  - Alignment with SSL representation in the latent space
  - Using Manifold structure

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

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

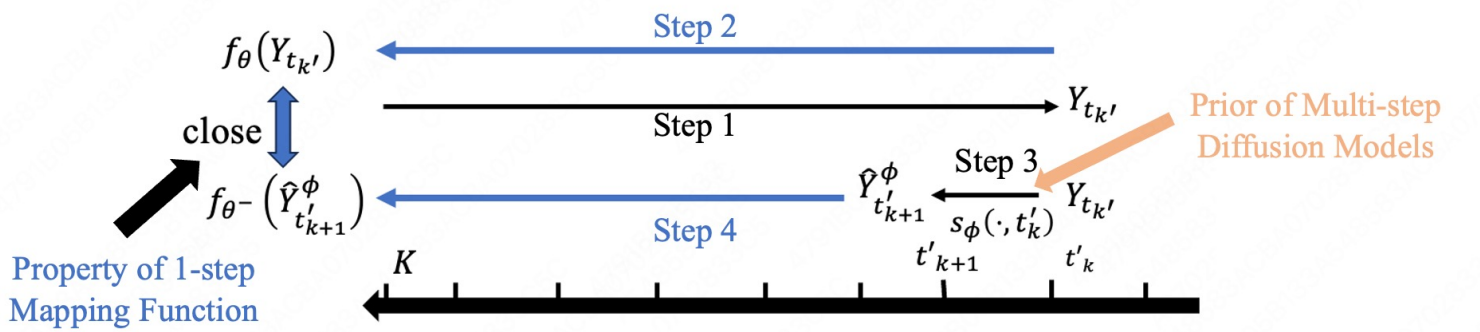
It is clear that the training of diffusion model do not involve label information.

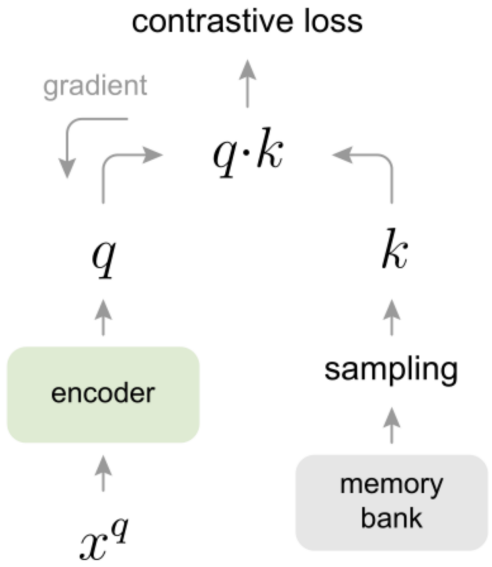
Consistency Distillation Objective Function[1]

- Goal: Use a NN $f_{\boldsymbol{\theta}}(Y_t, t)$ to approximate 1-step mapping function f

Let $\hat{Y}^{\phi}_{t_{k+1}}$ be the output running one step PFODE fromY_{t_k} with s_{ϕ}.

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

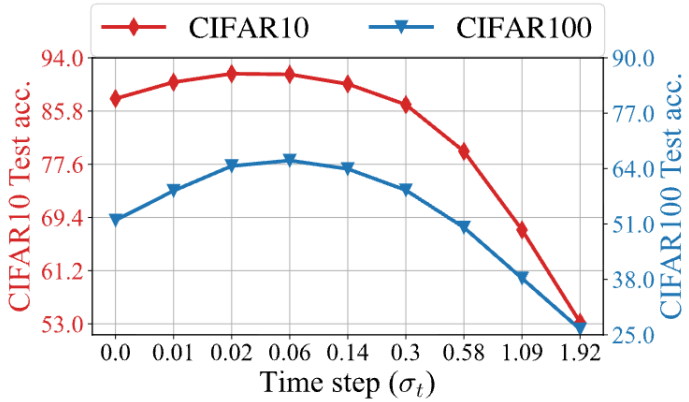




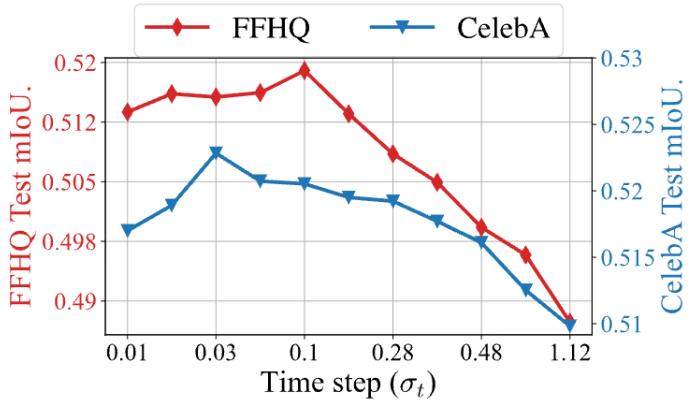
SSL MoCo Model

[1] SDCS, Consistency Models, ICML 2023.
[2] Momentum Contrast for Unsupervised Visual Representation Learning (MoCo). **Kaiming He**, Haoqi Fan, Yuxin Wu, **Saining Xie**, Ross Girshick

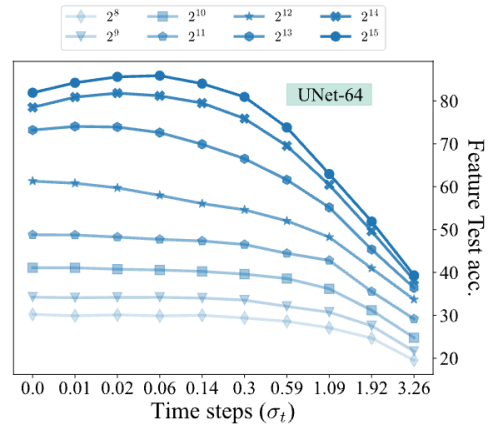
Diffusion Models can Play role of SSL Representation



(a) Classification



(b) Segmentation



- Diffusion model with great generalization property has unimodal dynamic for representation.
- Otherwise, suffers from a monotonically decreasing curve

[1] Understanding Representation Dynamics of Diffusion Models via Low-Dimensional Modeling

Diffusion Models can Play role of SSL Representation

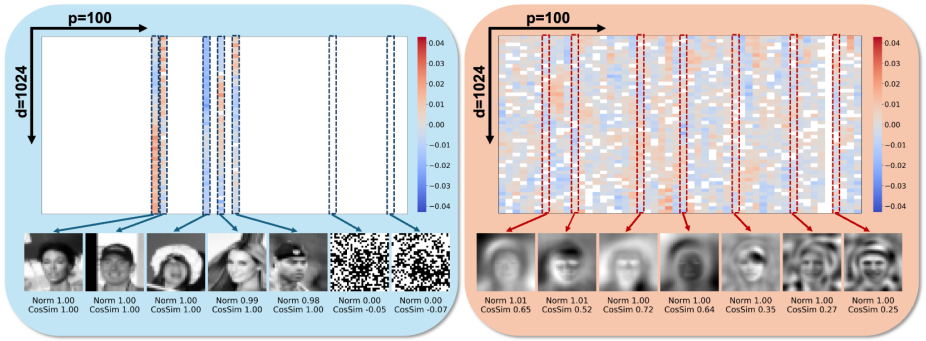
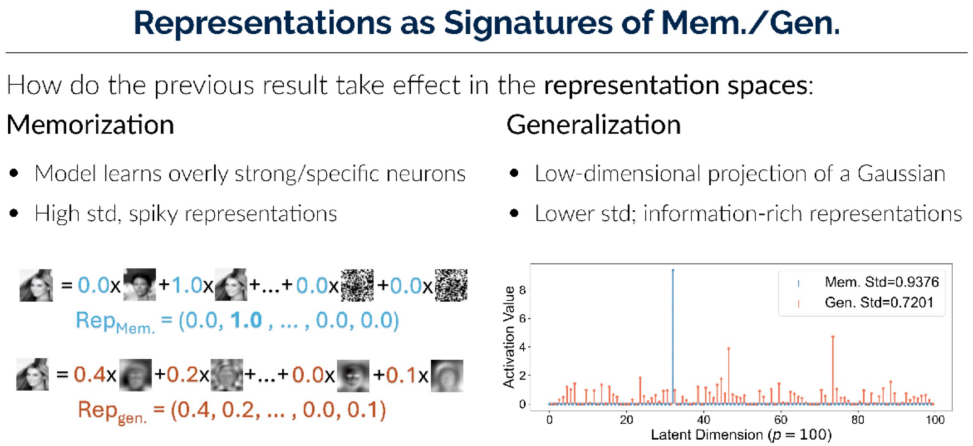


Figure 4: **Verification of Corollary 3.2 and Corollary 3.3.** We visualize the learned encoder matrix W_1 of a ReLU DAE trained with noise level $\sigma = 0.2$. When trained on 5 CelebA face images, the model stores training samples in its columns, matching Corollary 3.2. When trained on 10,000 images, the model generalizes and captures data statistics, consistent with Corollary 3.3. Empirically, the same behavior holds for larger noise, up to $\sigma = 5$; additional results are in Appendix A.1.



- Diffusion model with great generalization property Trends to learn a balanced property (similar to SSL methods)
- Otherwise, spiky representation

[1] Generalization of Diffusion Models Arises with a Balanced Representation Space

Overview

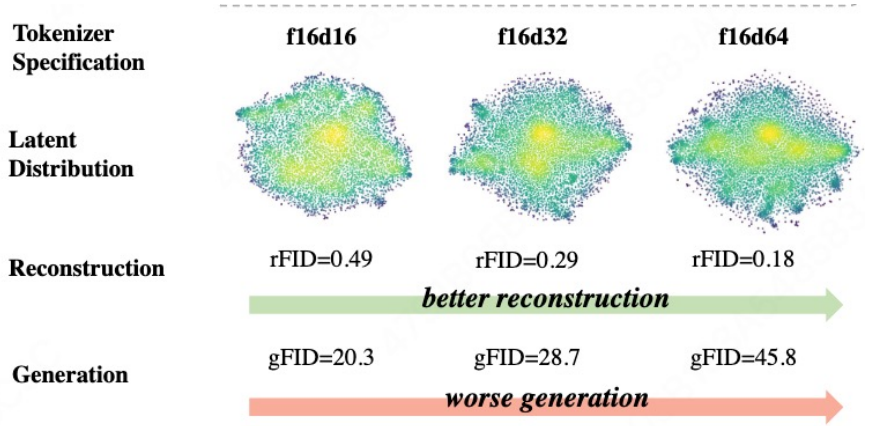
- The Relationship between Diffusion Models and SSL
- How to use Data Structure to improve Performance
  - Alignment with SSL representation in the latent space
  - Using Manifold structure

Alignment with SSL representation in the latent space

- Diffusion Models can Play the role of SSL Representation
- But not good enough compared with SSL method with large-size data
- Align with pretrained SSL models:
  - VAE level: VAVAE
  - Latent level: REPA and RAE

VAVAE: Constraint VAE Latent with Representation

- Due to the lack of semantic information constraints, VAEs struggle to achieve a balance between reconstruction and generation.



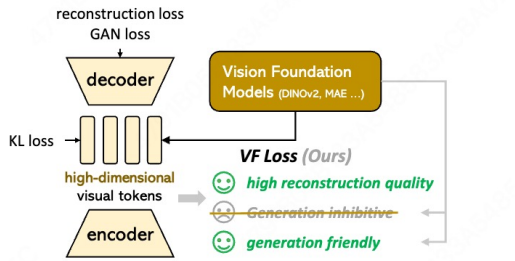
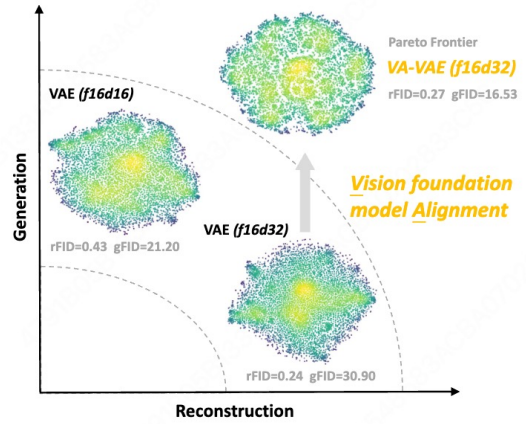
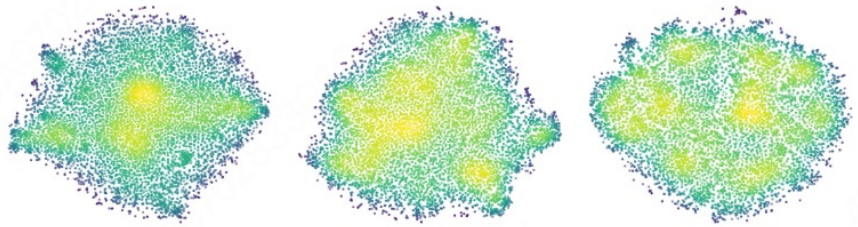


Figure 3. **The proposed Vision foundation model Aligned VAE (VA-VAE).** Vision foundation models are used to guide the training of high-dimensional visual tokenizers, effectively mitigating the optimization dilemma and improve generation performance.



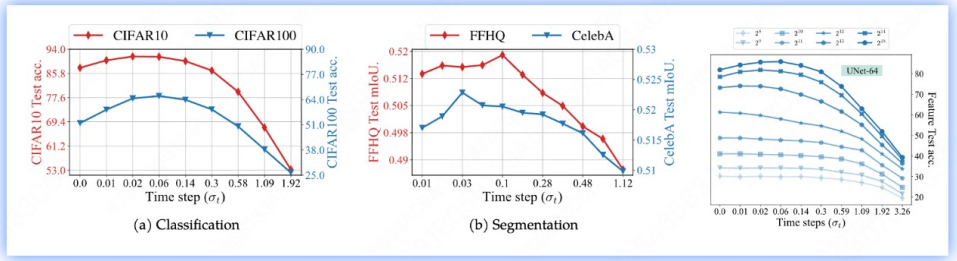


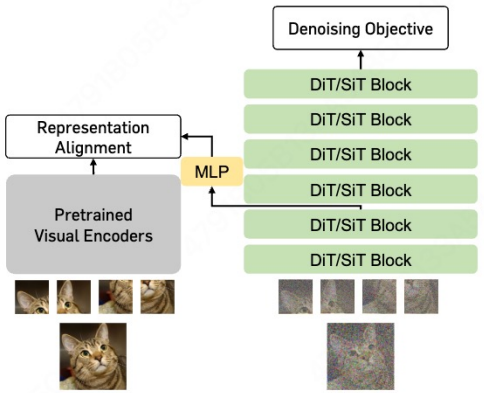
f16d32 naive | *f16d32* + VF MAE | *f16d32* + VF DINOv2

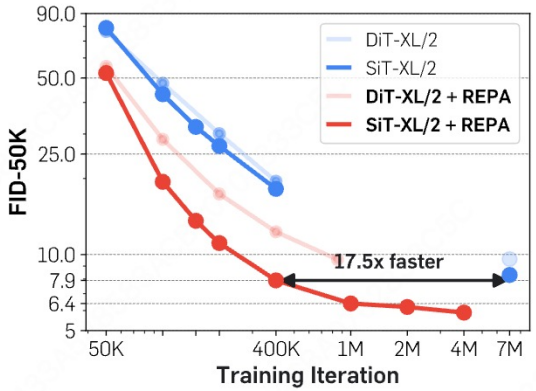
[1] Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion Models

Constraint Latent Diffusion with Representation

- Diffusion models tends to learn data representation
- (a) We directly align SSL with diffusion representation (REPA)
- Direct use SSL model as Encoder (RAE)







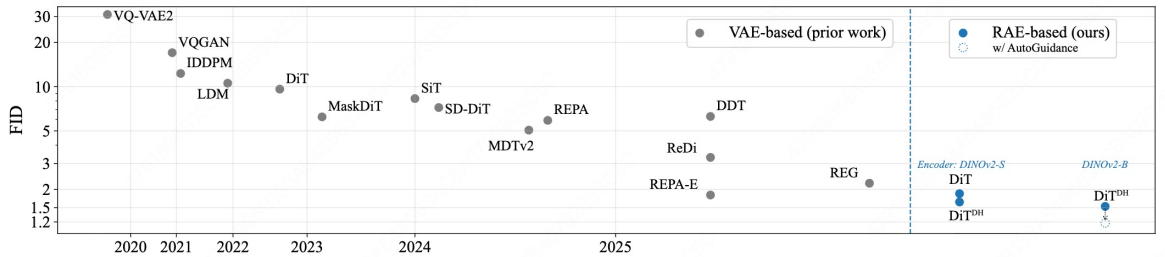


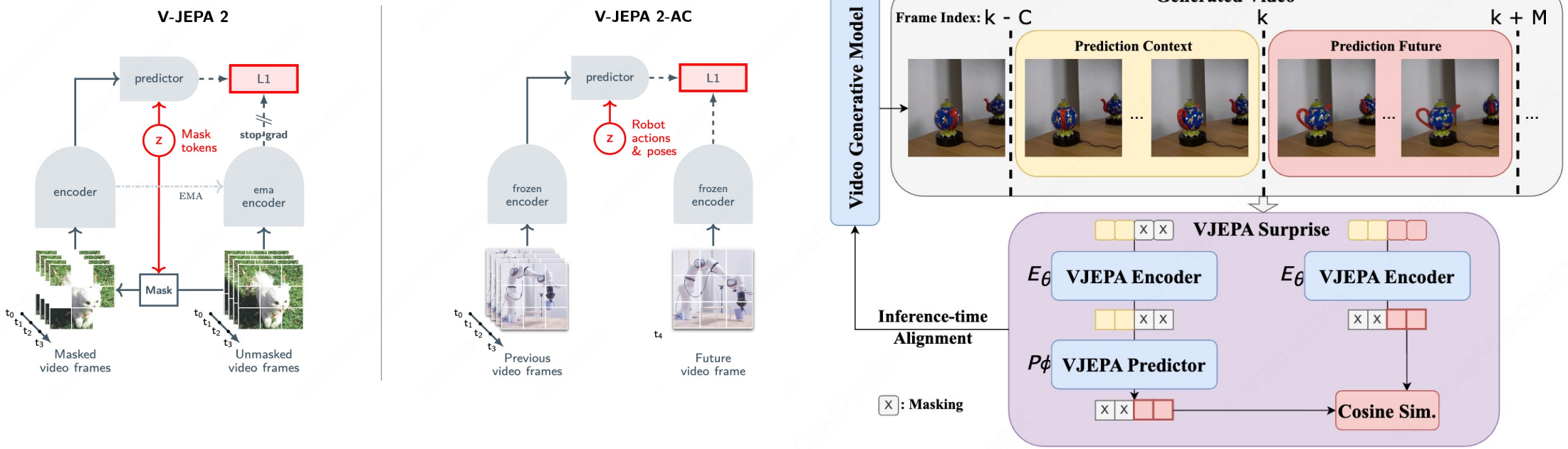
Figure 1: *Representation Autoencoder* (RAE) uses frozen pretrained representations as the encoder with a lightweight decoder to reconstruct input images without compression. RAE enables faster convergence and higher-quality samples in latent diffusion training compared to VAE-based models.

[1] REPRESENTATION ALIGNMENT FOR GENERATION:DIFFUSION TRANSFORMERSIS EASIER THAN YOU THINK

[2] DIFFUSION TRANSFORMERSWITH REPRESENTATION AUTOENCODERS

Guidance Inference with SSL Physical Information

- Due to the predict property, V-JEPA2 has physical information
- Use this information as reward guidance to enhance physical



[1] Inference-time Physics Alignment of Video Generative Models with Latent World Models
[2] V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning

Overview

- The Relationship between Diffusion Models and SSL
- How to use Data Structure to improve Performance
  - Alignment with SSL representation in the latent space
  - Using Manifold structure by predict clean samples

Predict Low-dim image instead of noise

- As the image has low dimension structure, directly predict image is easier compared to predict full space noise

$$\begin{cases} \boldsymbol{x}_\theta = \texttt{net}_\theta \\ \boldsymbol{z}_t = t\,\boldsymbol{x}_\theta + (1-t)\,\boldsymbol{\epsilon}_\theta \\ \boldsymbol{v}_\theta = \boldsymbol{x}_\theta - \boldsymbol{\epsilon}_\theta \end{cases}$$

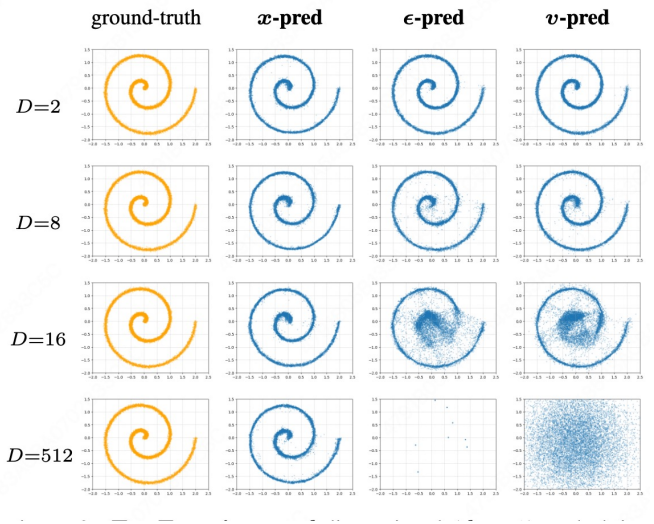
	(a) x-pred $\boldsymbol{x}_\theta := \texttt{net}_\theta(\boldsymbol{z}_t, t)$	(b) ϵ-pred $\boldsymbol{\epsilon}_\theta := \texttt{net}_\theta(\boldsymbol{z}_t, t)$	(c) v-pred $\boldsymbol{v}_\theta := \texttt{net}_\theta(\boldsymbol{z}_t, t)$
(1) x-loss: $\mathbb{E}\|\boldsymbol{x}_\theta - \boldsymbol{x}\|^2$	$\boldsymbol{x}_\theta$	$\boldsymbol{x}_\theta = (\boldsymbol{z}_t - (1-t)\boldsymbol{\epsilon}_\theta)/t$	$\boldsymbol{x}_\theta = (1-t)\boldsymbol{v}_\theta + \boldsymbol{z}_t$
(2) ϵ-loss: $\mathbb{E}\|\boldsymbol{\epsilon}_\theta - \boldsymbol{\epsilon}\|^2$	$\boldsymbol{\epsilon}_\theta = (\boldsymbol{z}_t - t\boldsymbol{x}_\theta)/(1-t)$	$\boldsymbol{\epsilon}_\theta$	$\boldsymbol{\epsilon}_\theta = \boldsymbol{z}_t - t\boldsymbol{v}_\theta$
(3) v-loss: $\mathbb{E}\|\boldsymbol{v}_\theta - \boldsymbol{v}\|^2$	$\boldsymbol{v}_\theta = (\boldsymbol{x}_\theta - \boldsymbol{z}_t)/(1-t)$	$\boldsymbol{v}_\theta = (\boldsymbol{z}_t - \boldsymbol{\epsilon}_\theta)/t$	$\boldsymbol{v}_\theta$

	x-pred	ϵ-pred	v-pred
x-loss	10.14	379.21	107.55
ϵ-loss	10.45	394.58	126.88
v-loss	8.62	372.38	96.53

(a) ImageNet 256×256, **JiT-B/16**

	x-pred	ϵ-pred	v-pred
x-loss	5.76	6.20	6.12
ϵ-loss	3.56	4.02	3.76
v-loss	3.55	3.63	3.46

(b) ImageNet 64×64, **JiT-B/4**



[1] Back to Basics: Let Denoising Generative Models Denoise (JiT)

Predict and perturb video to correct sample

- By predict clean samples to localize the sample to the manifold

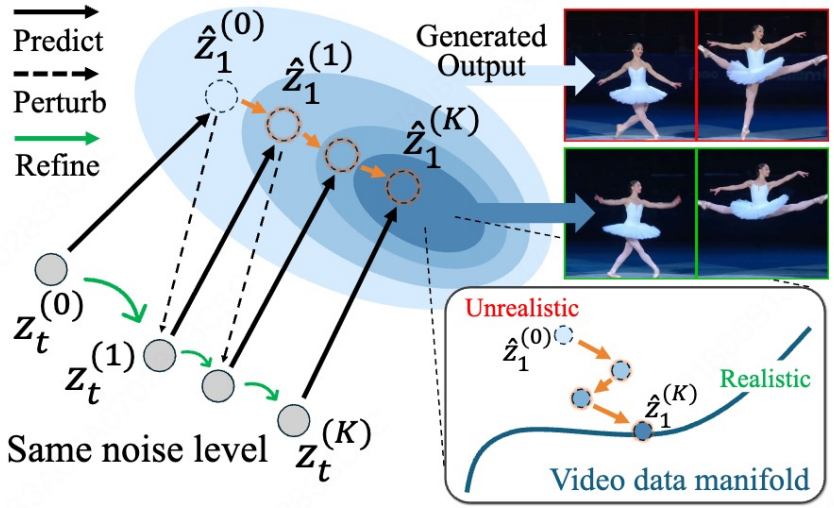
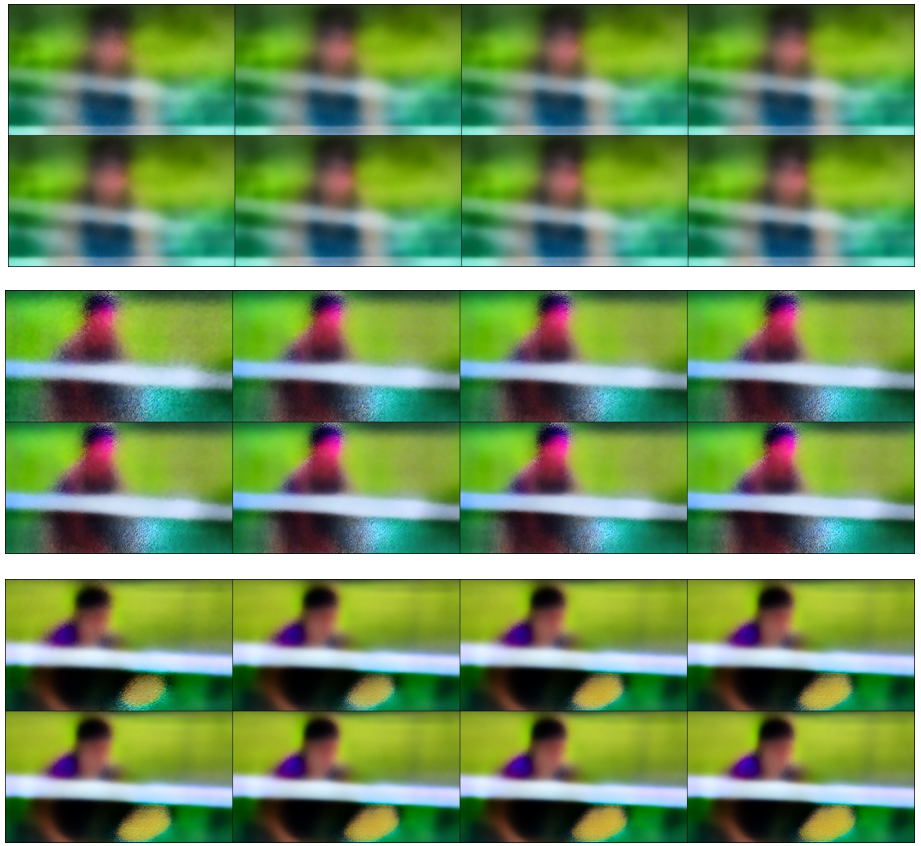


Figure 1. Concept of the **self-refining video sampling**. Within the same noise level, the video latent z_t is refined as the predicted endpoint $\hat{z}_1$ is pulled toward the data manifold.



[1] Self-Refining Video Sampling. 20260126.

Future work

- Current manifold learning mainly from intuition
- Can we model a more realistic manifold assumption?
- Can we use the manifold or data structure in the GRPO post training phase?

Future work

- Current manifold learning mainly from intuition
- Can we model a more realistic manifold assumption?
- Can we use the manifold or data structure in the GRPO post training phase?