

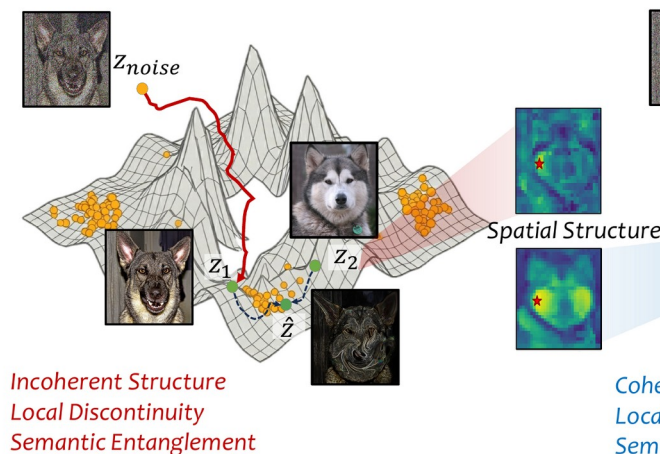
Prior-Aligned Autoencoders

What Matters for Diffusion-Friendly Latent Manifold?

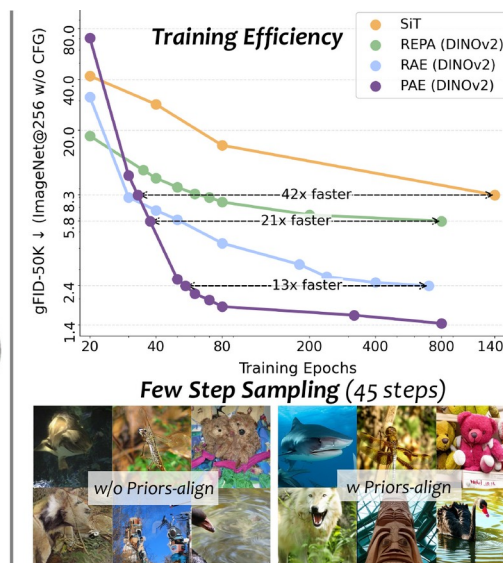
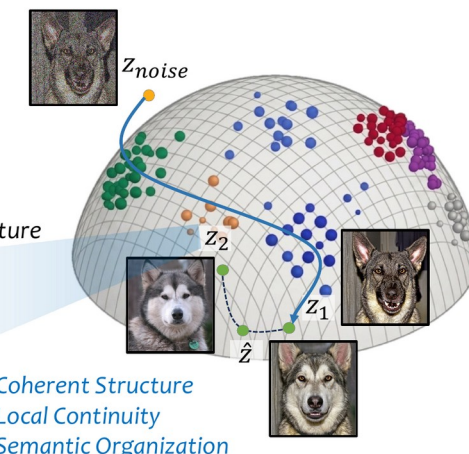
Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu,
Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, Yali Wang

Shanghai Jiao Tong University · Alibaba Group · SIAT, CAS · Beihang University
Sun Yat-sen University · Nankai University · Shanghai AI Laboratory

Reconstruction-oriented Latent Manifold



Priors-aligned Latent Manifold



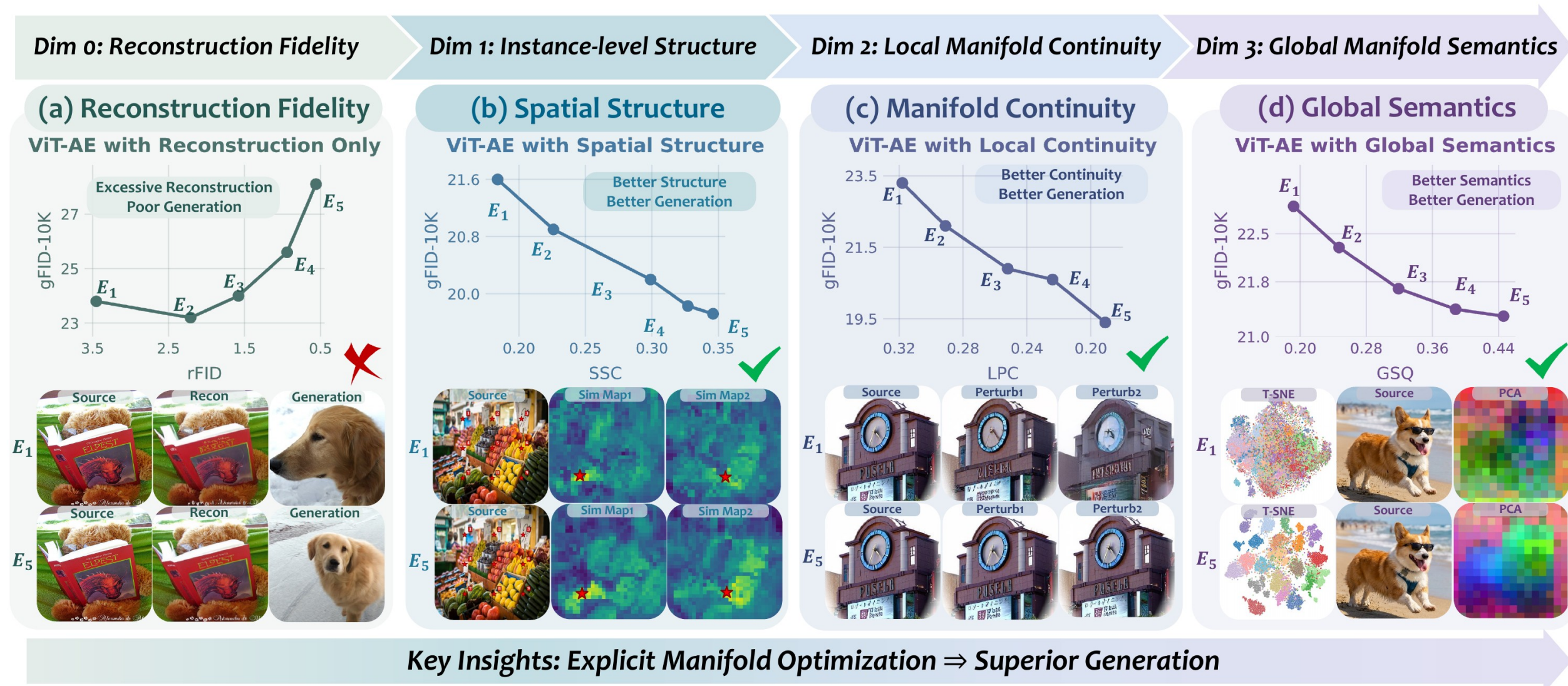
Paper



GitHub

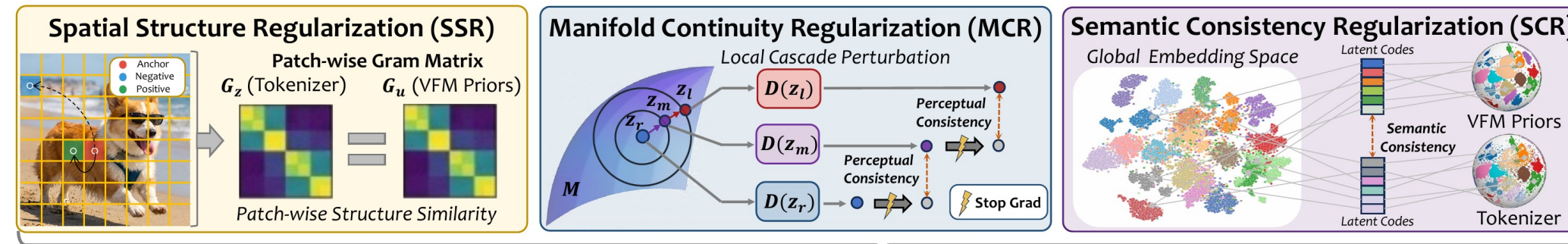
Reconstruction quality is an unreliable proxy for generation

Controlled tokenizer variants separate three diffusion-relevant manifold properties.



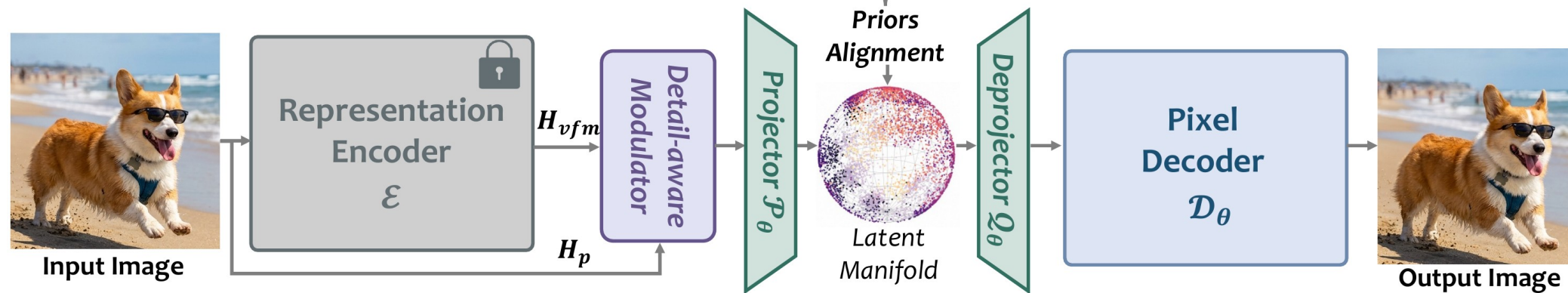
SSC $\uparrow$ Spatial structure LPC $\downarrow$ Local perceptual continuity GSQ $\uparrow$ Global semantics

PAE preserves VFM semantics while recovering pixel detail



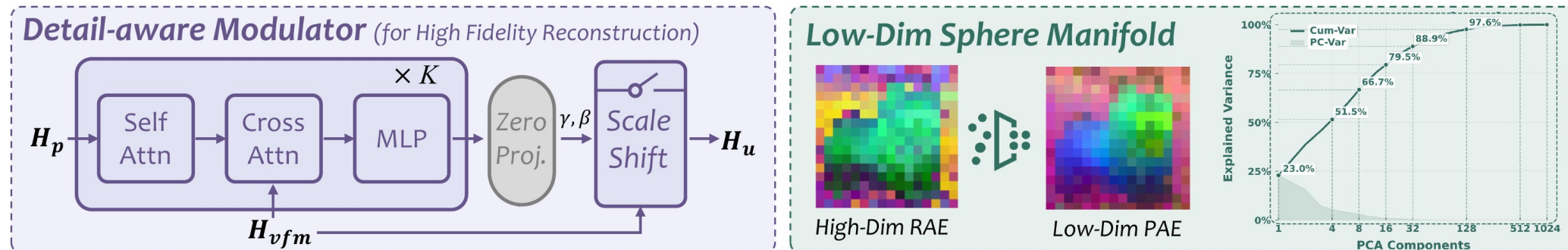
Frozen semantic reference

Keep the VFM encoder fixed during tokenizer training.



Detail-Aware Modulator

Use zero-initialized scale-and-shift fusion to inject image detail.



Encoder Tuning Free ☒

High Fidelity Reconstruction ☒

Dimension Redundancy Free ☒

Compact latent

16 × 16 × 32 latent
RMS normalization
6 DAM blocks

Trainable: DAM, projector, deprojector and pixel decoder | Frozen: VFM encoder

Three prior-alignment losses target complementary properties

SSR

Spatial structure

Align spatial Gram matrices with refined VFM targets.

Preserve relationships among spatial latent tokens.

MCR

Local continuity

Perturb the latent along one direction at two magnitudes.

Match decoded outputs from large $\rightarrow$ medium $\rightarrow$ anchor, using stop-gradient targets.

SCR

Global semantics

Align pooled and patch-token semantic directions after compression.

Match the semantic target to the compact bottleneck.

Training objective: reconstruction + weighted SSR + MCR + SCR

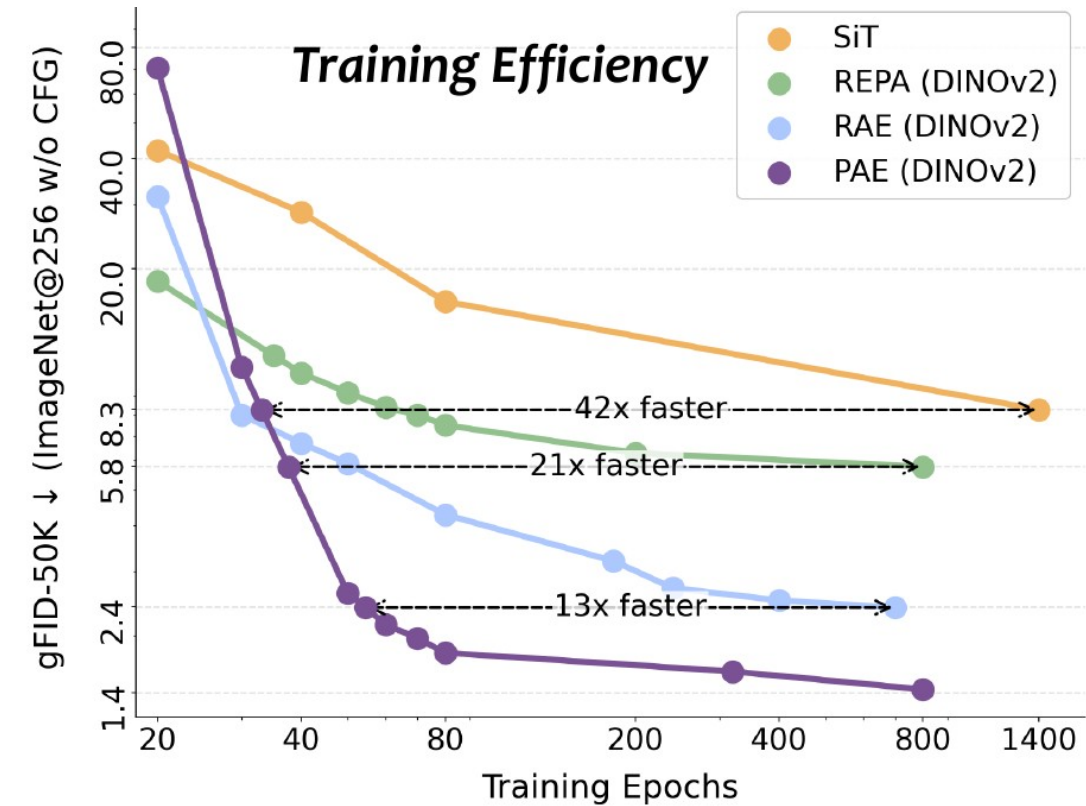
PAE improves generation quality and convergence

ImageNet 256 × 256 · LightningDiT-XL · guided gFID ↓

Method	Epochs	gFID ↓
REPA-E	800	1.15
FAE	800	1.29
VTP	600	1.11
PAE (DINOv2)	80	1.27
PAE (DINOv2)	800	1.03

1.03

guided gFID at 800 generator epochs



up to 13x

fewer epochs to comparable RAE quality

The prior objectives provide complementary gains

Objectives	gFID ↓	IS ↑
None	7.18	117.2
SSR only	2.74	161.8
MCR only	2.53	173.6
SCR only	2.63	168.4
SSR + MCR + SCR	1.86	210.8

25 tokenizer epochs
Same baseline and generator setup

Refined VFM targets

Target	gFID ↓
Raw	1.95
Refined	1.86

Target refinement helps, while the three prior losses account for the larger improvement.

Each objective most strongly improves its intended geometry dimension.

Tokenizers should organize latents for diffusion

Spatial coherence, local continuity and semantic organization are explicit targets for tokenizer design.



Evidence

1.03 guided gFID
1.27 at 80 epochs
Up to 13× fewer epochs

Scope

Evaluated on ImageNet 256×256 and class-conditional generation.

Paper



GitHub



arxiv.org/abs/2605.07915
github.com/ZhengrongYue/PAE