There is No VAE: End-to-End Pixel-Space Generative Modeling via Self-Supervised Pre-training

Jiachen Lei

Background

- **Current Paradigm:** Most high-resolution diffusion models rely on a compressed latent space provided by a pre-trained VAE.
- **The VAE Problem:**
  - Training VAEs is difficult and often produces imperfect reconstructions.
  - The VAE acts as a permanent performance bottleneck, limiting the generative model's capacity.
- **The Challenge:** Prior pixel-space models struggle with **high computational costs** and **slow convergence**.

Core Idea

- **Hypothesis:** The training of diffusion-based generative models can be decomposed into two stages, similar to classifier training in Self-Supervised Learning (SSL) field.
- **Role Decomposition:**
  - Encoder: Primarily learns high-level visual semantics from noisy inputs.
  - Decoder: Acts as a low-level pixel generator conditioned on those representations.

Training FrameWork

- **Two-Stage Training:**
  - Stage 1: Pre-train encoders to capture semantics from clean images while aligning them along deterministic ODE sampling trajectories.
  - Stage 2: Integrate the encoder with a randomly initialized decoder for end-to-end fine-tuning.

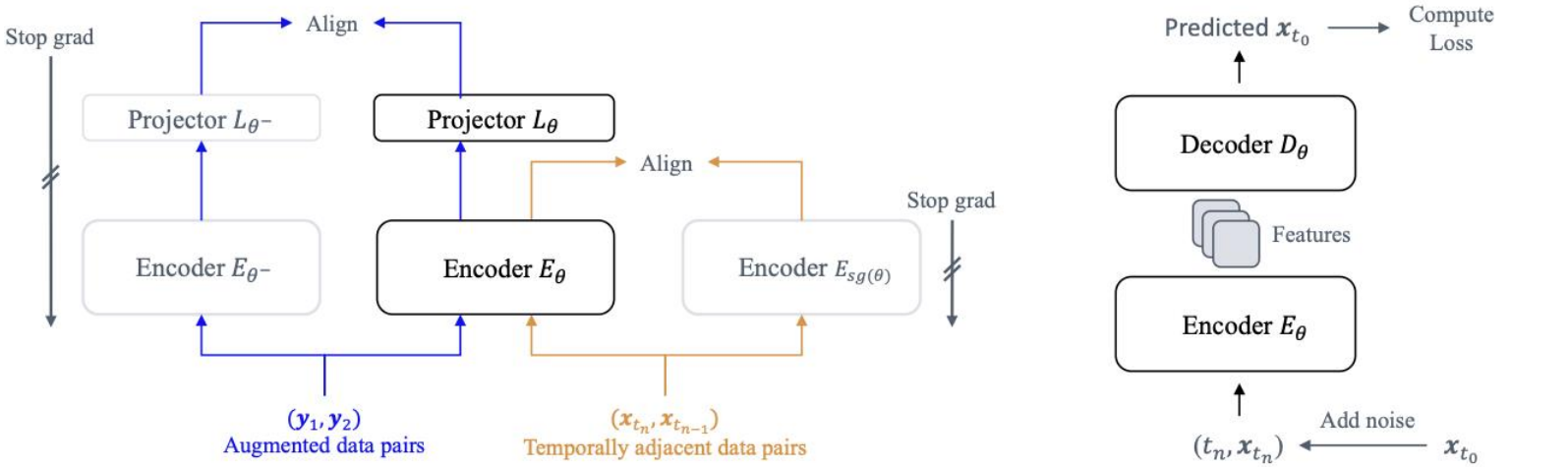

Figure 1: Training framework.

Main Results

Table 1: Diffusion model performance on ImageNet256

Model	FID↓	IS↑	Precision↑	Recall↑	NFE↓	Epochs	#Params	GFLOPs↓
Models in Latent Space								
LDM (Rombach et al., 2022)	3.60	247.7	**0.87**	0.48	250×2	167	55 + 400M	336 + 104
USP DiT-XL/2 (Chu et al., 2025)	2.33	267.0	-	-	250×2	240	84 + 675M	312 + 119
MaskDiT (Zheng et al., 2024)	2.28	276.6	0.80	0.61	79×2	1600	84 + 675M	312 + 119
DiT-XL/2 (Peebles & Xie, 2023)	2.27	278.2	0.83	0.57	250×2	1400	84 + 675M	312 + 119
SiT-XL/2 (Ma et al., 2024)	2.06	277.5	0.83	0.59	250×2	1400	84 + 675M	312 + 119
REPA (Yu et al., 2025)	1.42	305.7	0.80	0.65	434	800	84 + 675M	312 + 119
RAE (Zheng et al., 2025)	1.28	262.9	-	-	50×2	800	415 + 839M	107 + 146
Models in Pixel Space								
ADM† (Dhariwal & Nichol, 2021)	3.94	215.8	0.83	0.53	500	963	673M	761
RIN (Jabri et al., 2023)	3.42	182.0	-	-	1000	480	410M	334
SiD (Hoogeboom et al., 2023)	2.44	256.3	-	-	250×2	800	2.46B	555
VDM++ (Kingma & Gao, 2023)	2.12	278.1	-	-	250×2	-	2.46B	555
PixNerd-XL/16 (Wang et al., 2025)	1.93	297.0	0.79	0.59	100×2	320	700M	134
PixelFlow (Chen et al., 2025c)	1.98	282.1	0.81	0.60	-	-	677M	2909
SiD2 (Hoogeboom et al., 2025)	1.72	-	-	-	-	-	-	137
EPG-XL/16	2.04	283.2	0.80	0.61	75	800	583M	128
EPG-XXL/16	1.87	287.6	0.80	0.63	75	800	789M	176
EPG-G/16	1.70	297.8	0.80	0.63	75	1600	1391M	321
JiT-H/16 (Li & He, 2025)	1.86	**303.4**	0.78	0.62	191	600	953M	182
JiT-G/16 (Li & He, 2025)	1.82	292.6	0.79	0.62	191	600	2B	383
EPG-XXL/16	1.81	294.6	0.80	0.61	75	600	789M	176
EPG-G/16	1.75	275.1	0.80	0.62	75	600	1391M	321
EPG-G/16	**1.58**	298.4	0.80	0.63	75	1600	1391M	321

Table 2: Consistency model performance on ImageNet256.

Model	FID↓	NFE↓	Epochs	#Params
Models in Latent Space				
iCT-XL/2 (Song et al., 2023)	34.24	1	-	84M + 675M
	20.30	2	-	84M + 675M
Shortcut-XL/2 (Frans et al., 2025)	10.60	1	250	84M + 675M
	7.80	4	250	84M + 675M
IMM (Zhou et al., 2025)	8.05	1	6395	84M + 675M
Models in Pixel Space				
EPG-L/16	8.82	1	560	540M

Table 3: Training efficiency comparison with DiT.

Model	FID	Cost (hours)	#Params
sd-vae-mse†	-	160	84M
Our Pre-train	-	57	106M
DiT-XL/2†	2.27	506	675M
EPG-XL/16	2.04	139	583M
EPG-XXL/16	1.87	160	789M

Thank You*

*: Please refer to our paper for more details.