Why Move Beyond Vanilla VAEs to Representation Encoders?

Analysis: Why Direct Use of Representation Features Fails

PS-VAE: Making Representation Encoders Generation-Ready

The Platonic Representation Hypothesis

Neural networks trained on different data, objectives, and modalities converge to a shared representation of reality.

Visual Understanding

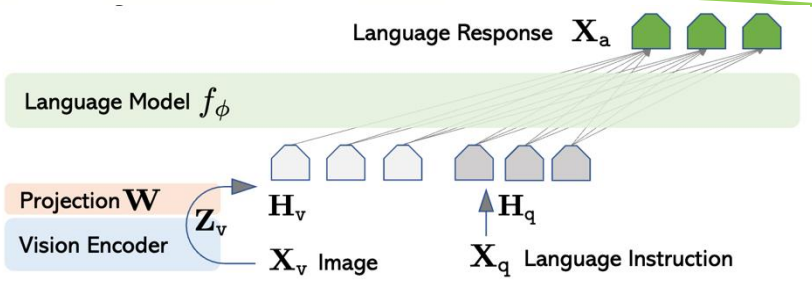


Figure 1: LLaVA network architecture.

Visual Representation Encoder

LLM

cat dog elephant

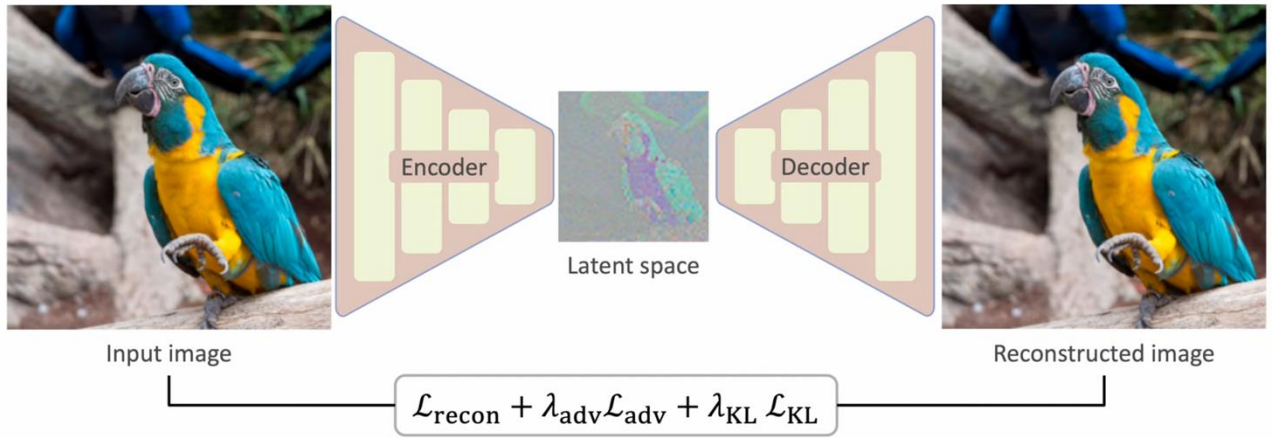
Visual Generation

Why?

VAE Space

Better VAE Representations Benefit Generation

Reconstruction under Information Bottleneck



Observation

⚡ EQVAE[1] → Tighter Information Bottleneck

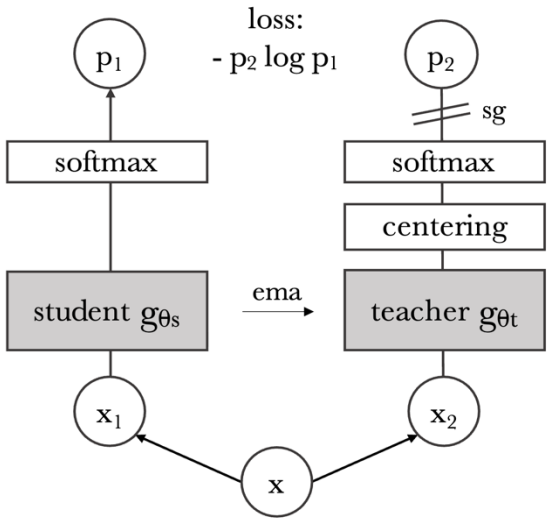
Observation

⚡ VAVAE[2] → Align to Representation encoder

Why not ?

Powerful Representation Learning Technique

DINO



iglip2

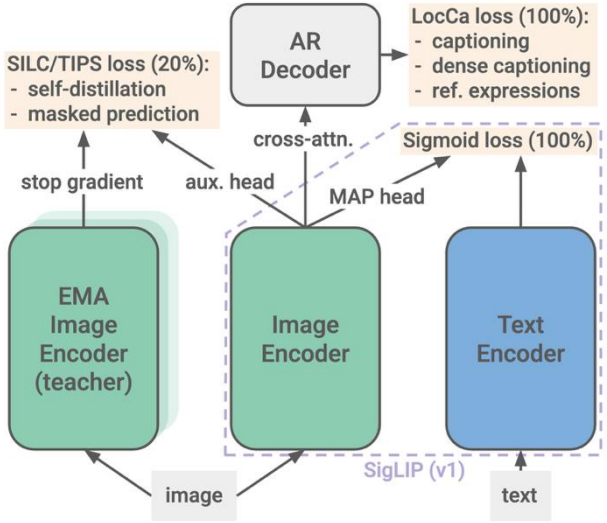


Figure adapted from LiteVAE: Lightweight and Efficient Variational Autoencoders for Latent Diffusion Models.
1.EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling
2. Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion Models

Why Move Beyond Vanilla VAEs to Representation Encoders?

Analysis: Why Direct Use of Representation Features Fails

PS-VAE: Making Representation Encoders Generation-Ready

Visualization comparison between RAE and VAE.

Resolution: 256 × 256 Model size: ~700M Training data: CC12M

(a) Reconstruction Comparison

Remove the *elderly man* wearing glasses

(b) Editing Comparison

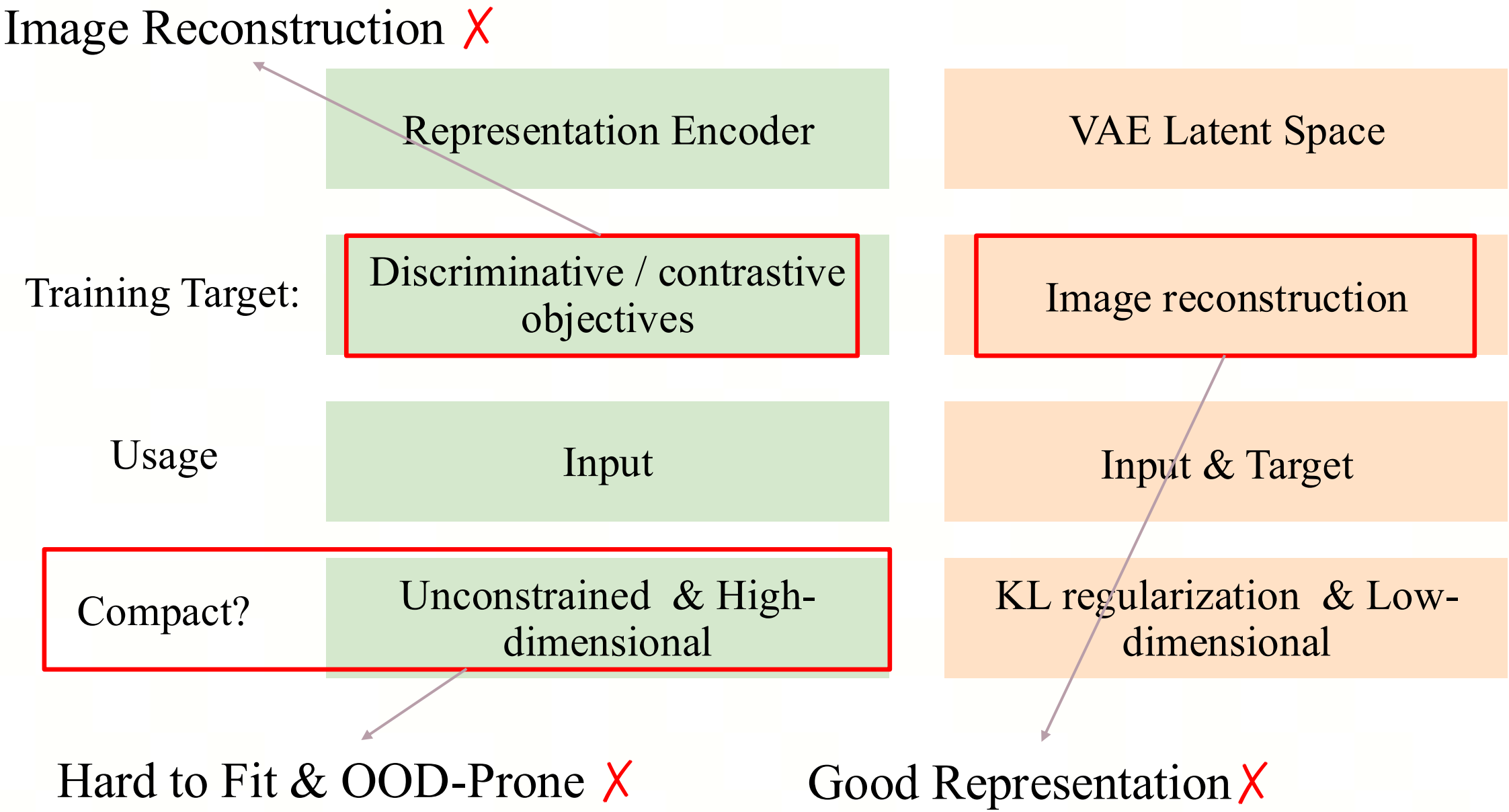
A traditional Venetian mask with intricate designs and feather embellishments is …

Two square-shaped pink erasers rest on …

Two snowboards are positioned side by side on a …

(c) Text to Image Comparison

Representation Encoder vs. VAE Latent Space



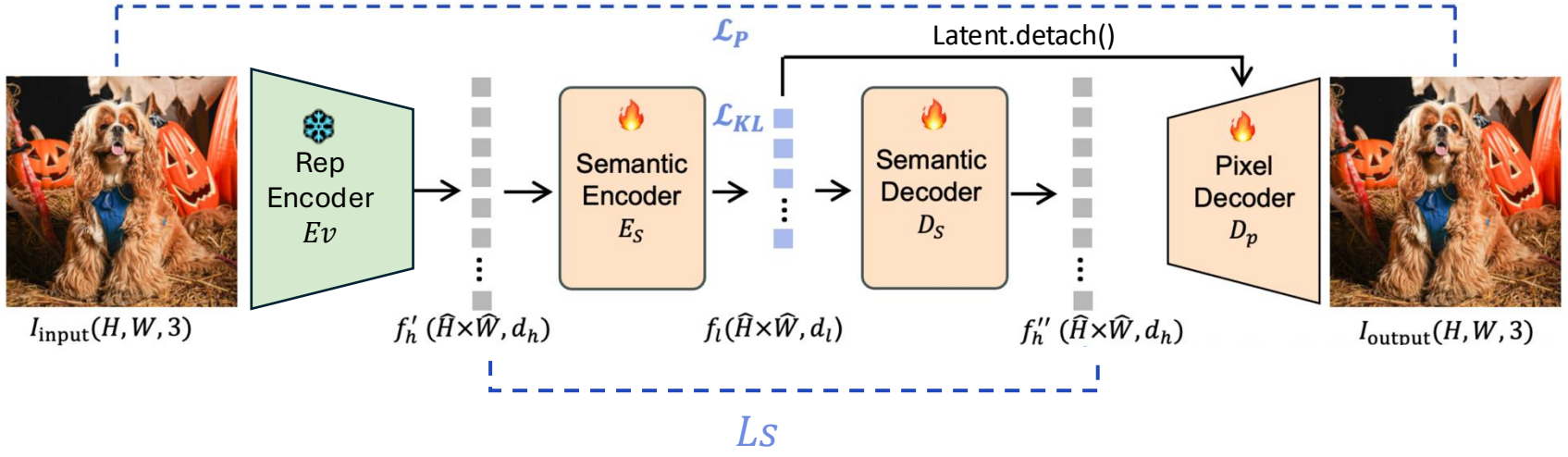
Why Move Beyond Vanilla VAEs to Representation Encoders?

Analysis: Why Direct Use of Representation Features Fails

PS-VAE: Making Representation Encoders Generation-Ready

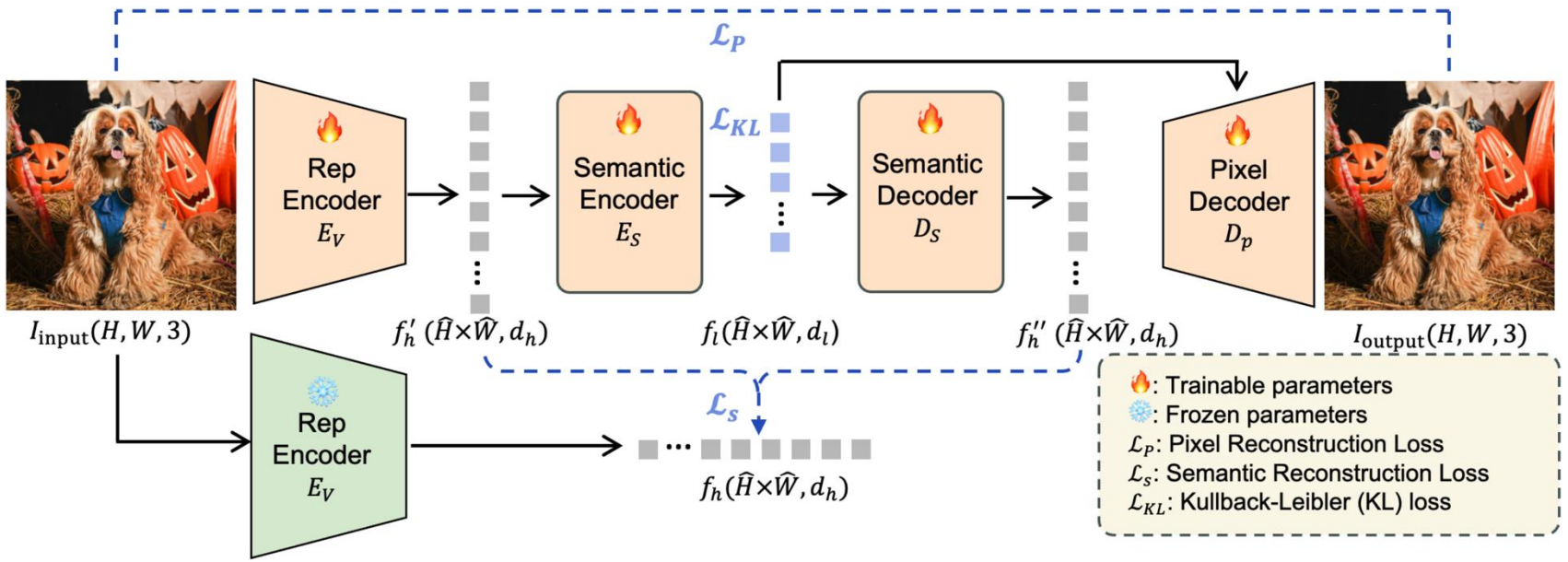
PS-VAE: Semantic Stage

🔥: Trainable parameters
❄️: Frozen parameters
$\mathcal{L}_P$: Pixel Reconstruction Loss
$\mathcal{L}_s$: Semantic Reconstruction Loss
$\mathcal{L}_{KL}$: Kullback-Leibler (KL) loss



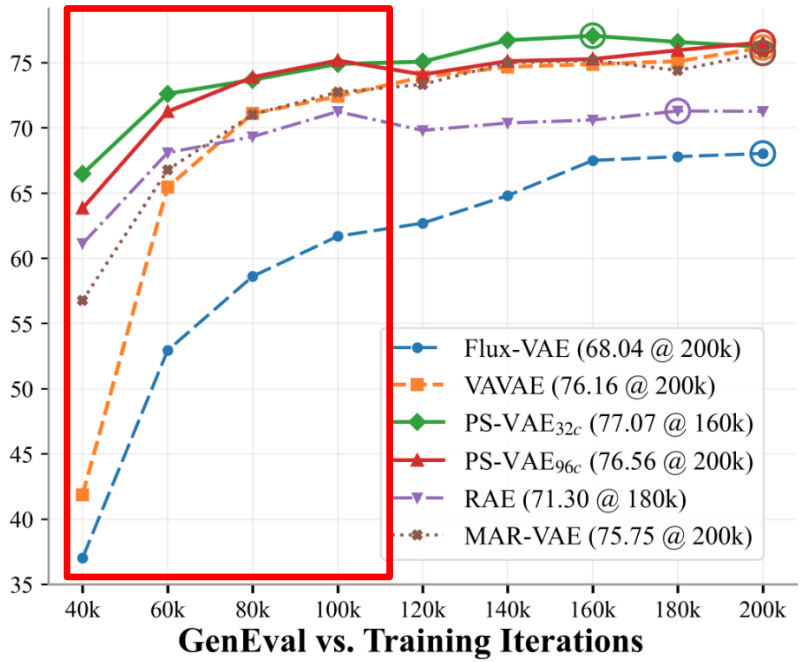
Method	rFID (↓)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	Linear Probe (↑)	GenEval (↑)	DPG (↑)	Editing Reward (↑)
MAR-VAE	0.534	26.18	0.135	0.715	5.4/15.1	75.7	83.2	0.06
RAE	0.619	19.20	0.254	0.436	83.0/96.6	71.3	81.7	0.06
S-VAE	1.407	17.78	0.296	0.390	81.1/95.7	73.7	83.6	0.12

PS-VAE: Pixel-Semantic Stage

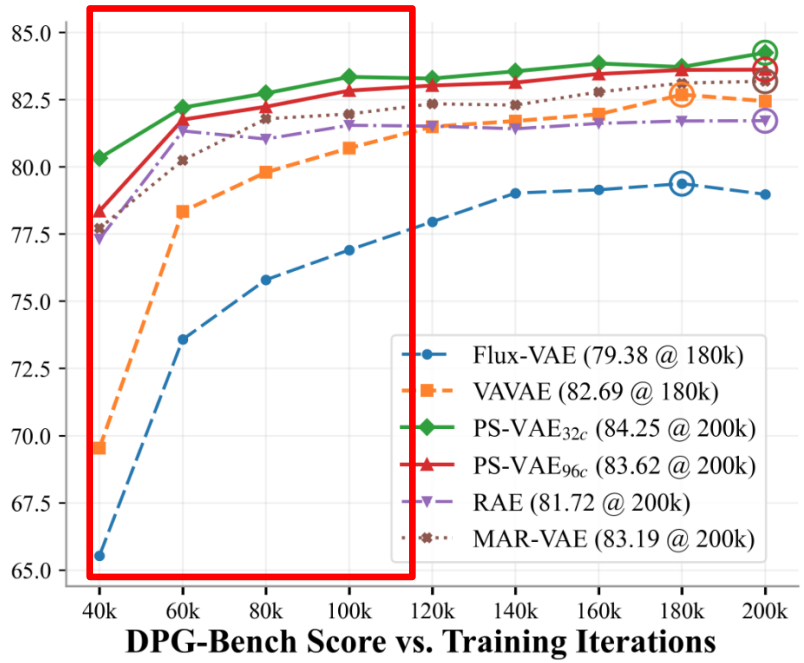


Method	rFID (↓)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	Linear Probe (↑)	GenEval (↑)	DPG (↑)	Editing Reward (↑)
MAR-VAE	0.534	26.18	0.135	0.715	5.4/15.1	75.7	83.2	0.06
RAE	0.619	19.20	0.254	0.436	83.0/96.6	71.3	81.7	0.06
S-VAE	1.407	17.78	0.296	0.390	81.1/95.7	73.7	83.6	0.12
P-VAE	0.398	29.81	0.073	0.850	11.8/26.7	75.2	82.1	0.04
PS-VAE	0.203	28.79	0.085	0.817	79.5/94.8	76.6	83.6	0.22

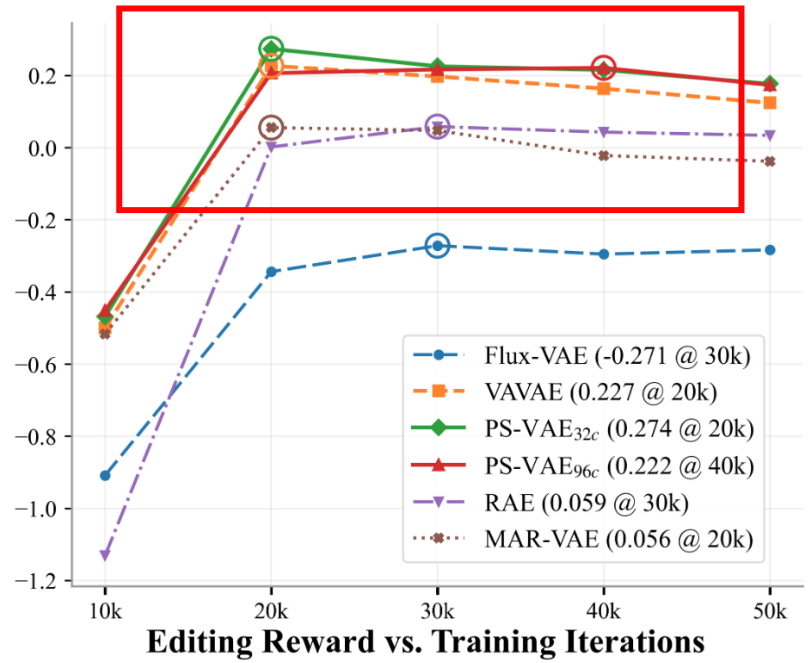
Method	rFID (↓)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	GenEval (↑)	DPG-Bench (↑)	Editing Reward (↑)
Flux-VAE (Labs, 2024)	**0.175**	**32.86**	**0.044**	**0.912**	68.04	78.98	-0.271
MAR-VAE (Li et al., 2024)	0.534	26.18	0.135	0.715	75.75	83.19	0.056
VAVAE (Yao et al., 2025)	0.279	27.71	0.097	0.779	76.16	82.45	<u>0.227</u>
RAE (Zheng et al., 2025)	0.619	19.20	0.254	0.436	71.27	81.72	0.059
PS-VAE$_{32c}$	0.584	24.53	0.168	0.662	<u>76.22</u>	**84.25**	**0.274**
PS-VAE$_{96c}$	<u>0.203</u>	<u>28.79</u>	<u>0.085</u>	<u>0.817</u>	**76.56**	<u>83.62</u>	0.222



GenEval vs. Training Iterations



DPG-Bench Score vs. Training Iterations



Editing Reward vs. Training Iterations

Thanks