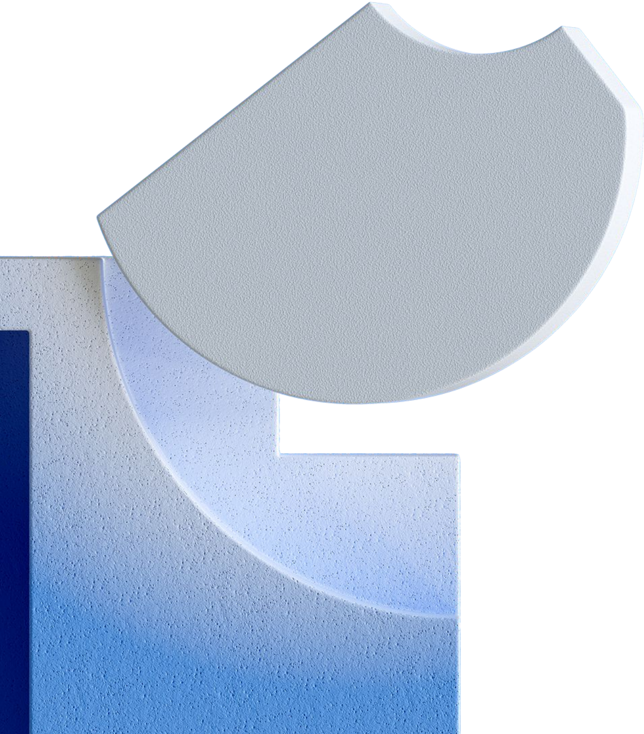
Google DeepMind



Epsilon-VAE: Denoising as Visual Decoding

Long Zhao, Sanghyun Woo, Ziyu Wan, Yandong Li, Han Zhang, Boqing Gong, Hartwig Adam, Xuhui Jia, Ting Liu

Google DeepMind

ICML 2025

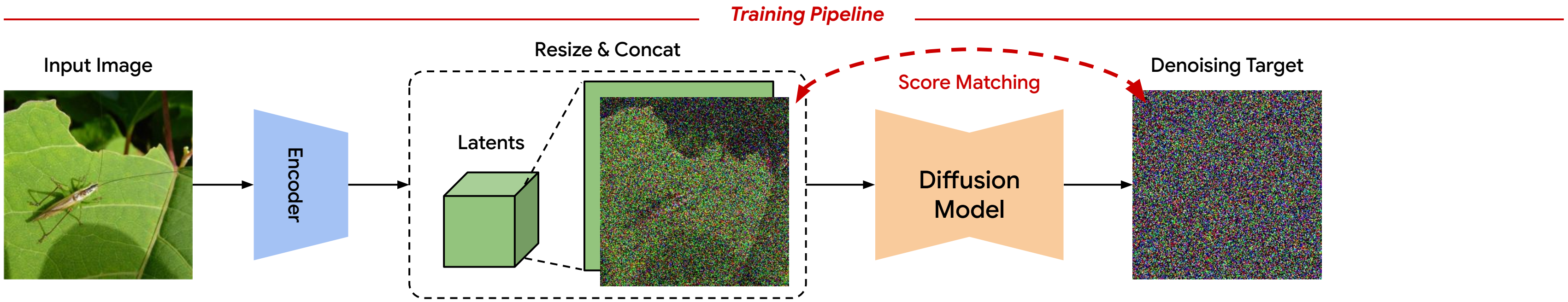
Google DeepMind

What is Epsilon-VAE?

A visual **autoencoder** where the decoder is replaced with a **diffusion process**, achieving better reconstruction performance than state-of-the-art VAEs.

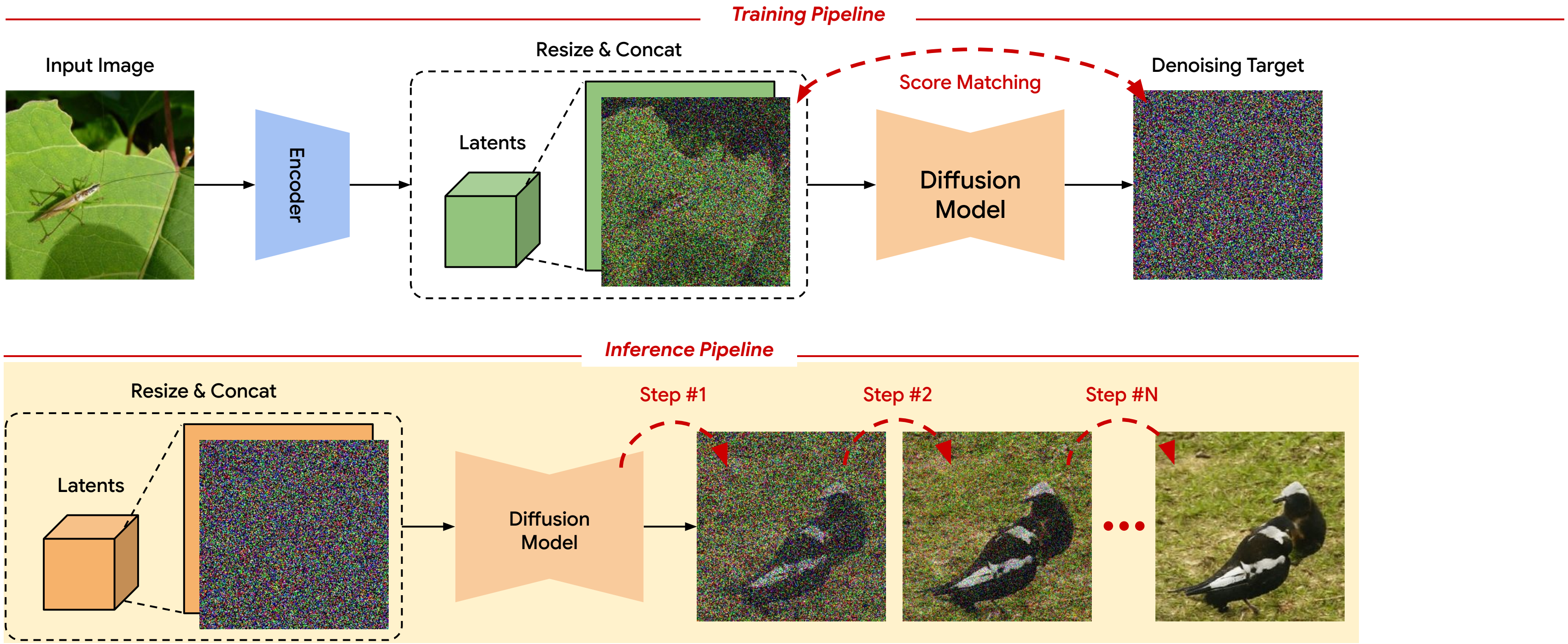
A visual autoencoder (or tokenizer) is essential for generative models: discrete tokens allow step-by-step conditional generation in autoregressive models, while continuous latents enable efficient learning in the denoising process of diffusion models.

Key problems & design



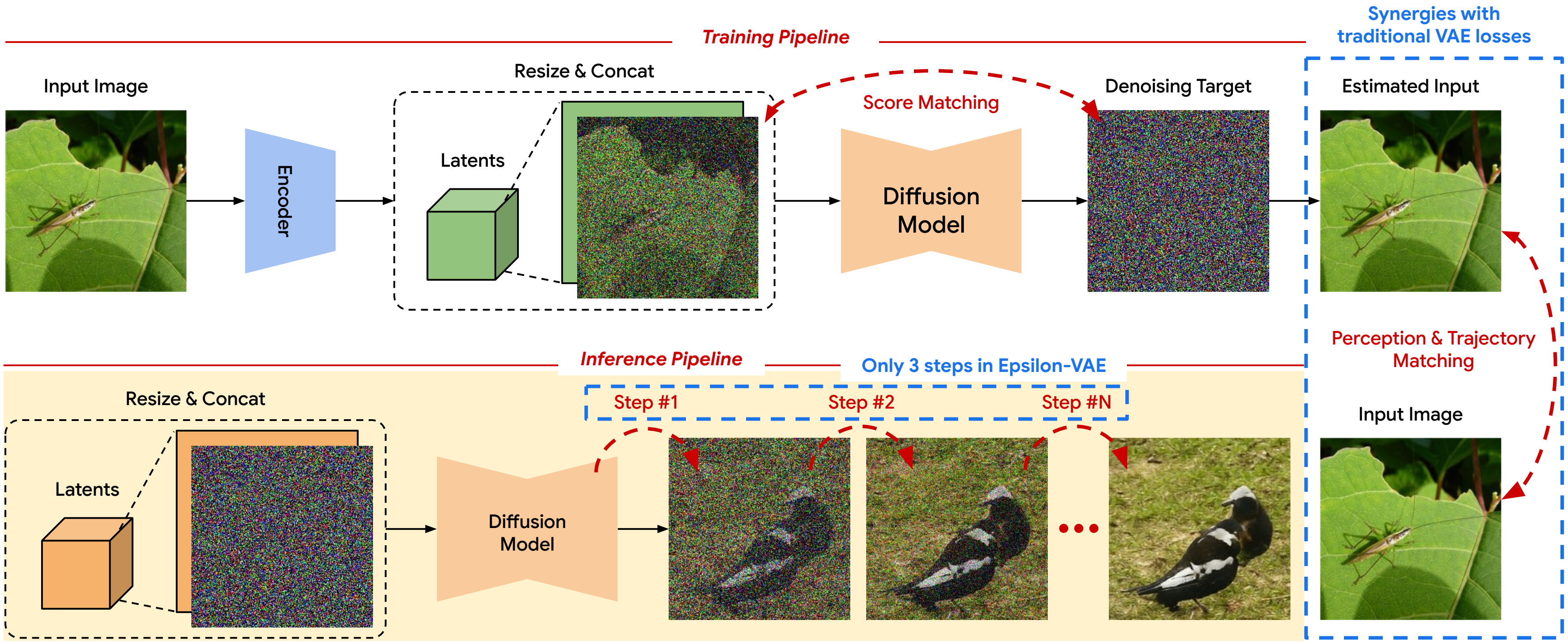
An overview of Epsilon-VAE. We frame visual decoding as an iterative denoising problem by replacing the autoencoder decoder with a diffusion model, optimized using a score matching losses.

Key problems & design



An overview of Epsilon-VAE. We frame visual decoding as an iterative denoising problem by replacing the autoencoder decoder with a diffusion model, optimized using a score matching losses. **During inference, images are reconstructed (or generated) from encoded (or sampled) latents through an iterative denoising process.**

Key problems & design



An overview of Epsilon-VAE. We frame visual decoding as an iterative denoising problem by replacing the autoencoder decoder with a diffusion model, optimized using **a combination of score, perception, and trajectory matching losses.** During inference, images are reconstructed (or generated) from encoded (or sampled) latents through an iterative denoising process. **The number of sampling steps N can be flexibly adjusted within small NFE regimes (from 1 to 3).**

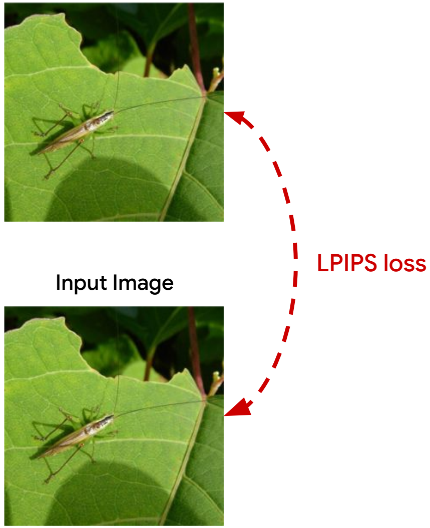
Loss functions

Perceptual matching

- We compute the LPIPS loss between reconstruction estimated by the model at time t (using the simple reversing step) the target real image.

Estimated x_0

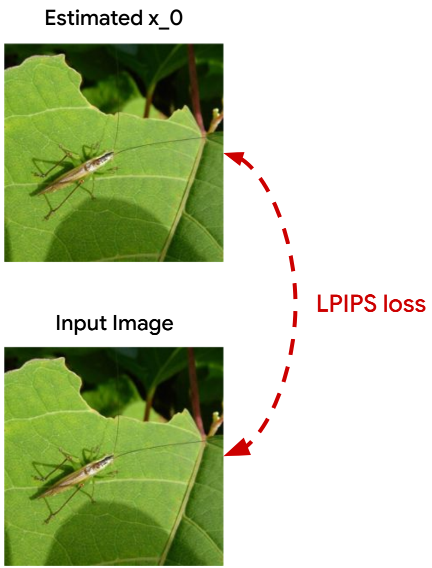
Input Image



Loss functions

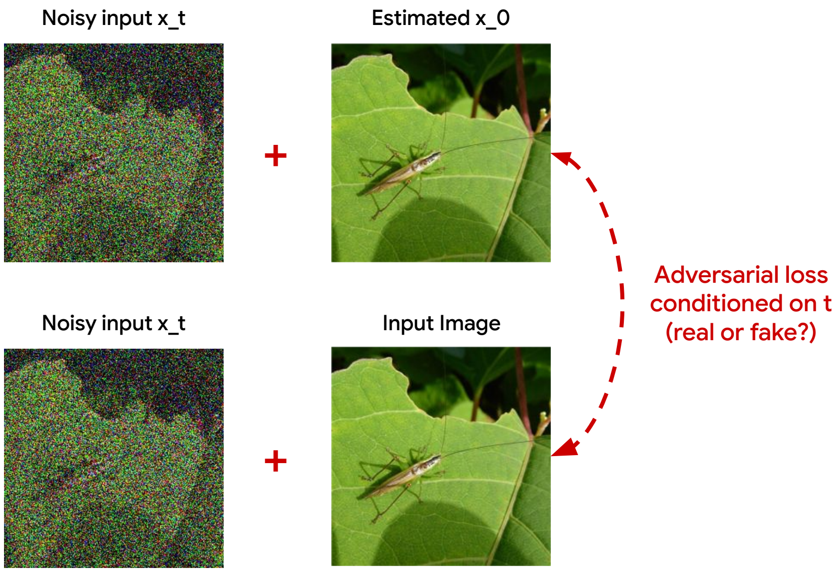
Perceptual matching

- We compute the LPIPS loss between reconstruction estimated by the model at time t (using the simple reversing step) the target real image.



Denoising trajectory matching

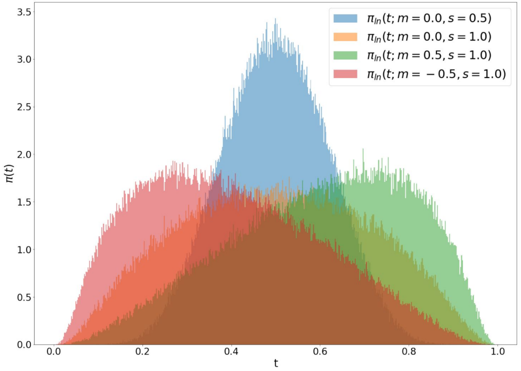
- We adapt the standard adversarial loss to enforce trajectory consistency from x_t to (estimated) x_0 rather than solely on estimated x_0.



Noise and time scheduling

Training

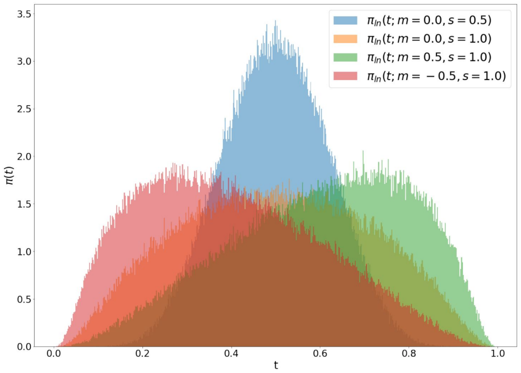
- We adopt the **rectified flow** parameterization.
- Noise scheduling can also be adjusted by scaling the intermediate states x_t with a constant fact, which shifts the signal-to-noise ratio downward. We scale x_t by 0.6 when we reconstruct 128 x 128 images, which makes training more challenging over time while preserving the shape of the trajectory (Chen, 2023).
- We sample t from a **logit-normal distribution**, which emphasizes intermediate timesteps (Esser et al., 2024).



Noise and time scheduling

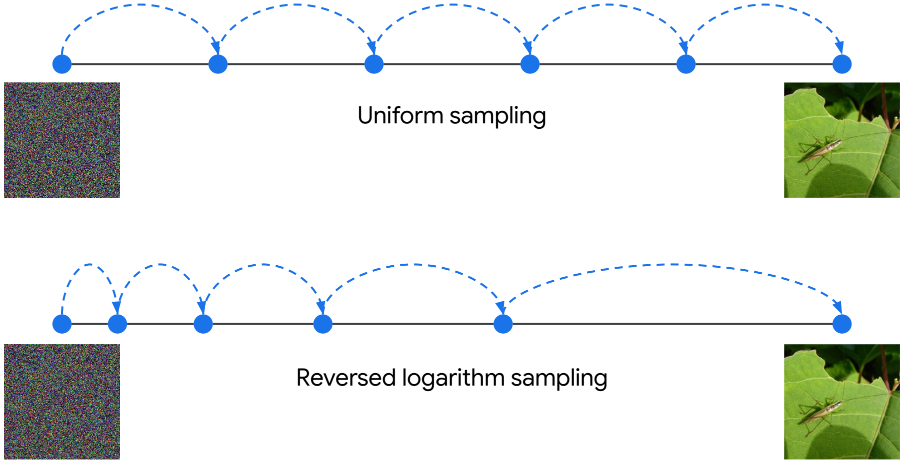
Training

- Noise scheduling can also be adjusted by scaling the intermediate states x_t with a constant fact, which shifts the signal-to-noise ratio downward. We scale x_t by 0.6 on reconstructing 128 x 128 images, which makes training more challenging over time while preserving the shape of the trajectory (Chen, 2023)
- We adopt the **rectified flow** parameterization.
- We sample t from a **logit-normal distribution**, which emphasizes intermediate timesteps (Esser et al., 2024).



Inference

- During sampling, we apply a reversed logarithm mapping, resulting in denser sampling steps early in the inference process.



Evaluation: Reconstruction quality

ImageNet reconstruction results (rFID) at different resolutions using VAEs trained at 128 × 128 under **Epsilon-VAE-SD** setup. * denotes training at 128 × 128 followed by fine-tuning at a higher resolution.

Method	IN 128 x 128 rFID	IN 256 x 256 rFID	IN 512 x 512 rFID	IN 256 x 256 rFID *
SD-VAE	4.54	1.21	0.91	0.86
LiteVAE	4.40	0.97	-	0.73
Epsilon-VAE (B)	1.94	0.65	0.61	0.52
Epsilon-VAE (M)	1.58	0.55	0.53	0.47
Epsilon-VAE (L)	1.47	0.52	0.41	0.45
Epsilon-VAE (XL)	1.34	0.49	0.39	0.43
Epsilon-VAE (H)	1.00	0.44	0.35	0.38

Key observations

- **Epsilon-VAE** effectively generalizes to higher resolutions, consistently preserving its performance advantage over other VAEs.
- Furthermore, we find that fine-tuning models at the target (higher) resolution leads to improvement at it.
- We hence utilize this **multi-stage training strategy** in the following experiments when the target resolution is larger than 128 × 128.

Evaluation: Reconstruction quality

Comparisons with state-of-the-art image autoencoders under **Epsilon-VAE-SD** setup. All results are computed on 256 × 256 ImageNet 50K validation set and COCO-2017 5K validation set. Epsilon-VAE-SD (M) achieves better reconstruction quality while having similar parameters (49M) in the decoder with other VAEs. Epsilon-VAE-SD (H) has 355M decoder parameters.

Downsample	Method	Latent dim.	ImageNet rFID	COCO-2017 rFID
16 x 16	VQGAN	256 (discrete)	5.74	3.69
	LlamaGen	8 (discrete)	4.63	2.69
	SD-VAE	4	4.78	2.78
	Epsilon-VAE (M)	4	4.42	2.41
	Epsilon-VAE (H)	4	4.29	2.37
8 x 8	VQGAN	4 (discrete)	3.90	2.06
	SD-VAE	4	2.79	2.02
	LiteVAE	4	2.60	1.92
	Epsilon-VAE (M)	4	2.38	1.82
	Epsilon-VAE (H)	4	2.31	1.78

Key observations

- **Epsilon-VAE** outperforms state-of-the-art VAEs when the decoder sizes are comparable, and its performance can be further improved by scaling up the decoder.

Evaluation: Ablation studies

Ablation study on key design choices for the Epsilon-VAE diffusion decoder under **Epsilon-VAE-lite** setup. A systematic evaluation of the proposed architecture [A], objectives [O], and noise & time scheduling [S]. Each row progressively modifies or builds upon the baseline decoder, showing improvements in performance.

Ablation	NFE	rFID
Baseline: DDPM-based diffusion decoder	1000	28.22
[O] (a) Diffusion → Rectified flow parameterization	100	24.11
[S] (b) Uniform → Logit-normal time step sampling during training	50	23.44
[A] (c) DDPM UNet → ADM UNet	50	22.04
[O] (d) Perceptual matching	10	11.76
[O] (e) Adversarial denoising trajectory matching	5	8.24
[S] (f) Scale diffusion inputs by 0.6	5	7.08
[S] (g) Uniform → Reversed logarithm time spacing during inference	3	6.24

Key observations

- In (a), Transitioning from standard diffusion to rectified flow (Liu et al., 2023) straightens the optimization path, resulting in significant gains in rFID and NFE.
- In (d), LPIPS loss is applied to match reconstructions with real images, leading to remarkable improvements.
- In (e), adversarial trajectory matching loss improve model understanding of the underlying optimization trajectory, significantly enhancing rFID scores and NFE.

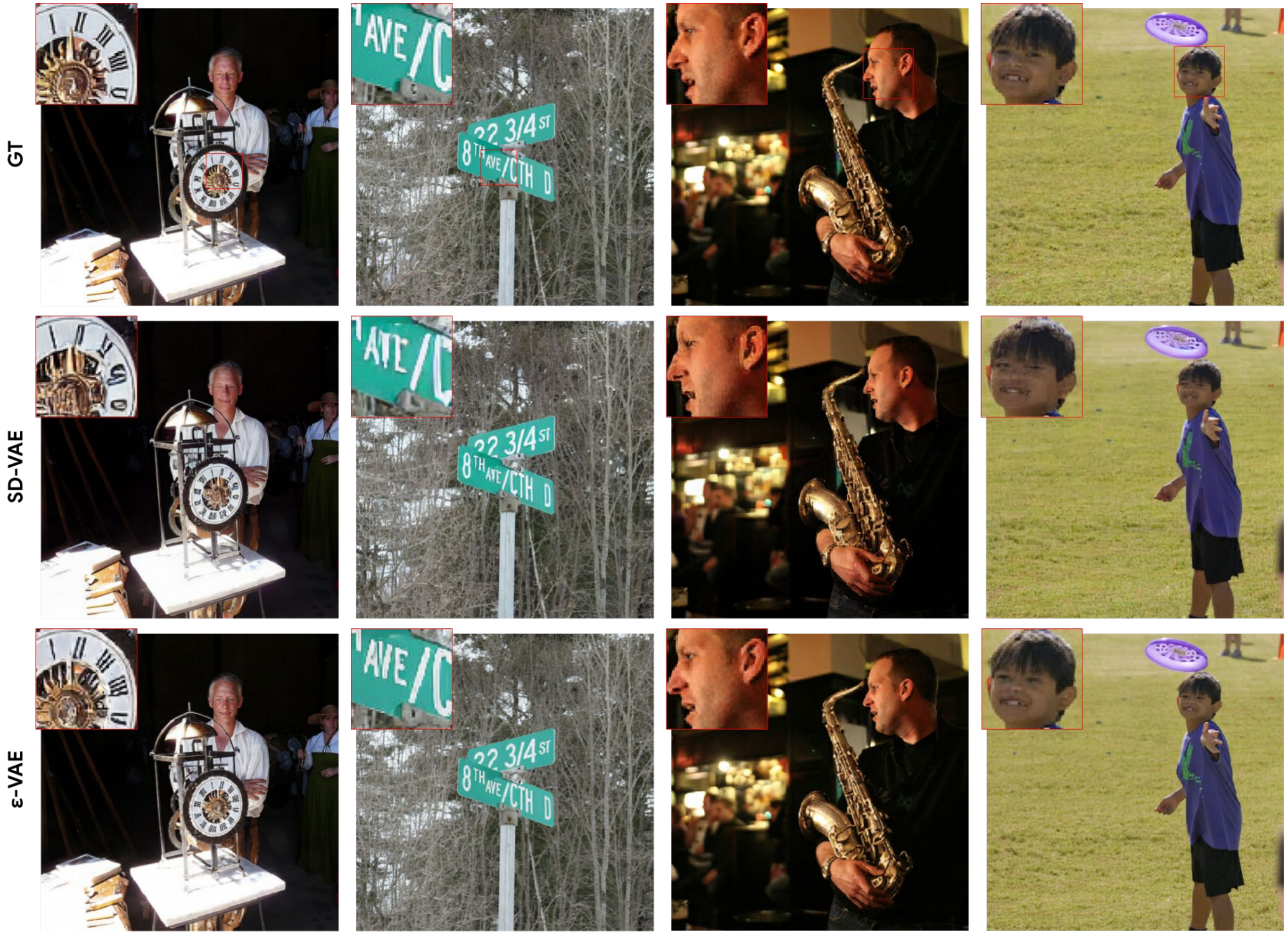


Image reconstruction results under the SD-VAE configuration (Rombach et al., 2022) at the resolution of 512 × 512.

Key observations

- We find that **Epsilon-VAE** produces more accurate visual details than **SD-VAE** in the highlighted regions with text or human face.

Evaluation: Conditional image generation

Benchmarking class-conditional image generation on ImageNet 256 × 256 under **Epsilon-VAE-SD** setup. We use the DiT-XL/2 architecture (Esser et al., 2024) for latent diffusion models, and we do not apply classifier-free guidance (Ho & Salimans, 2022).

Downsample	Method	Throughput (image/sec)	FID
16 x 16	SD-VAE	1220	14.59
	Epsilon-VAE (M)	1192	10.68
	Epsilon-VAE (H)	1180	9.72
8 x 8	Asym-VAE	502	10.85
	Omni-VAE	480	12.25
	SD-VAE	522	11.63
	Epsilon-VAE (M)	491	9.39
	Epsilon-VAE (H)	477	8.85

2.3x

Key observations

- **Epsilon-VAE** consistently outperforms other VAEs across different dowmsample factors.
- **Epsilon-VAE** achieves favorable generation quality while using only 25% of the token length typically required by **SD-VAE**.
- This token length reduction significantly accelerates latent diffusion model generation, leading to 2.3x higher inference throughput while maintaining competitive generation quality.

Thank you.