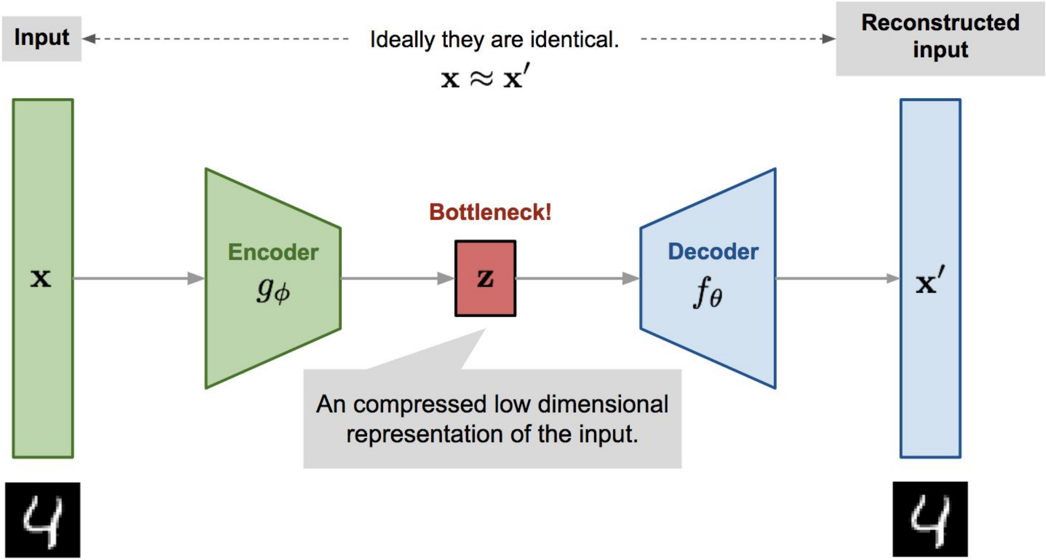
Stable Diffusion path

presented by
Vitaly Bondar
Generative DL enthusiast
johngull @ gmail

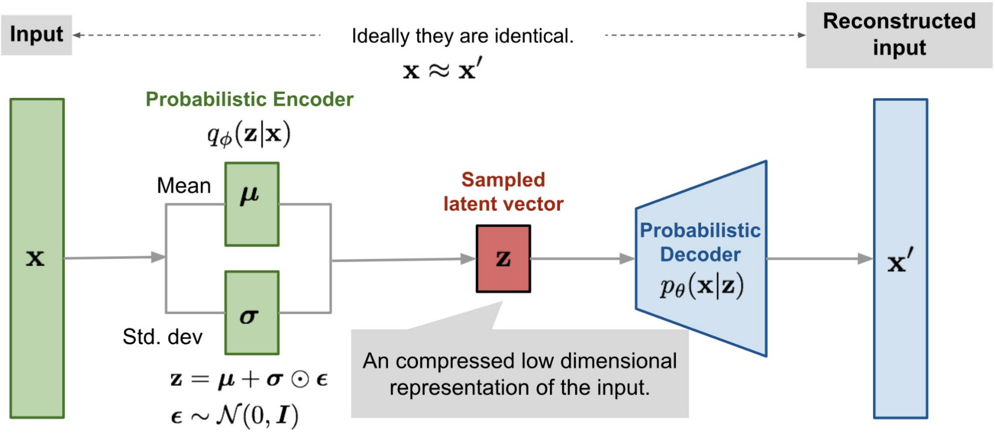
Autoencoder

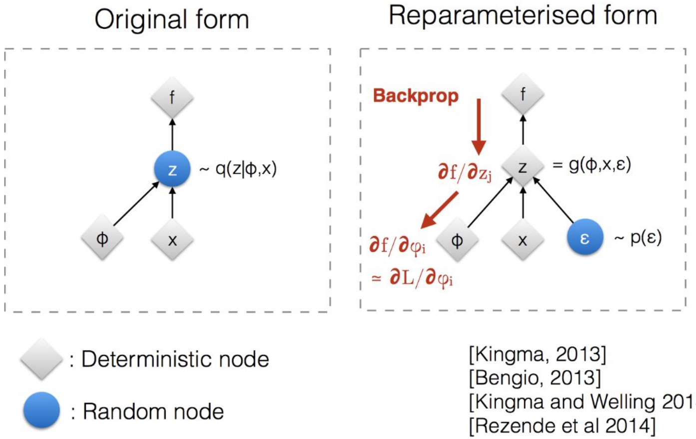


LeCun, Y. Modeles connexionistes de l'apprentissage ` . Ph.D. thesis, 1987

Illustration: Lilian Weng

Variational autoencoder (VAE)





[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

$$-L_{\text{VAE}} = \log p_\theta(\mathbf{x}) - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})) \leq \log p_\theta(\mathbf{x})$$

Kingma, Welling. Auto-Encoding Variational Bayes. 2013

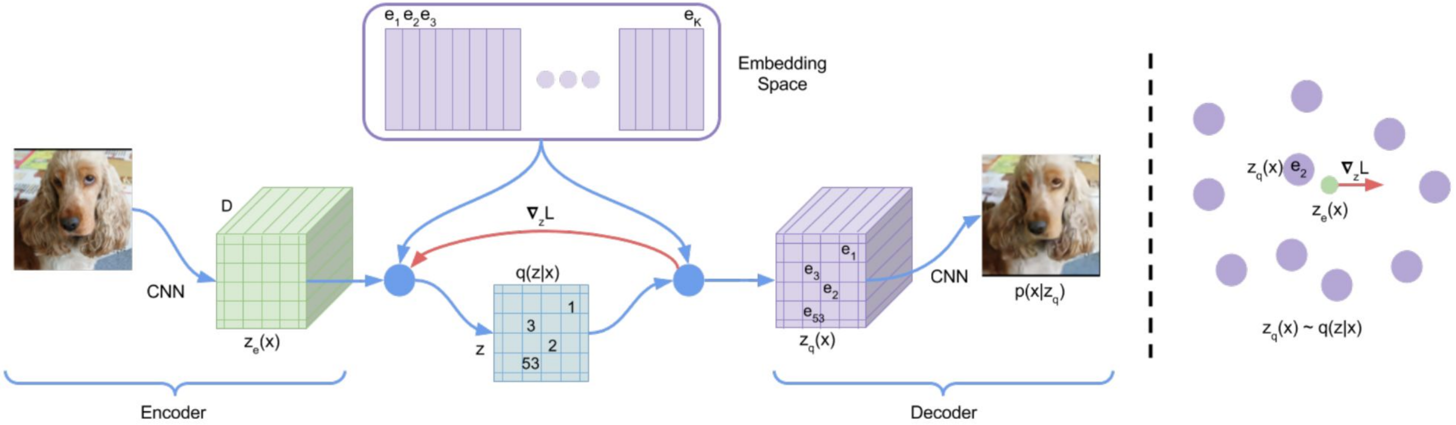
Illustrations:
Lilian Weng, D. Kingma

Variational autoencoder (VAE)

(a) Learned Frey Face manifold

Kingma, Welling. Auto-Encoding Variational Bayes. 2013

Vector Quantized Variational Autoencoder (VQ-VAE)



$$L = \log p(x|z_q(x)) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2,$$

Oord et al. Neural Discrete Representation Learning, 2017

Vector Quantized Variational Autoencoder (VQ-VAE)

Figure 2: Left: ImageNet 128x128x3 images, right: reconstructions from a VQ-VAE with a 32x32x1 latent space, with K=512.

Oord et al. Neural Discrete Representation Learning, 2017

Vector Quantized Variational Autoencoder (VQ-VAE)

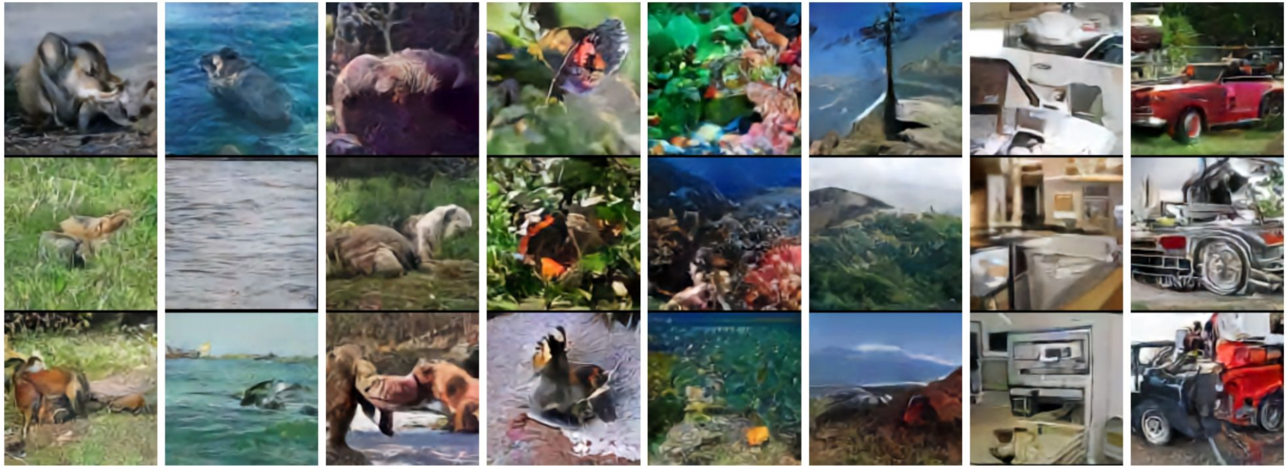
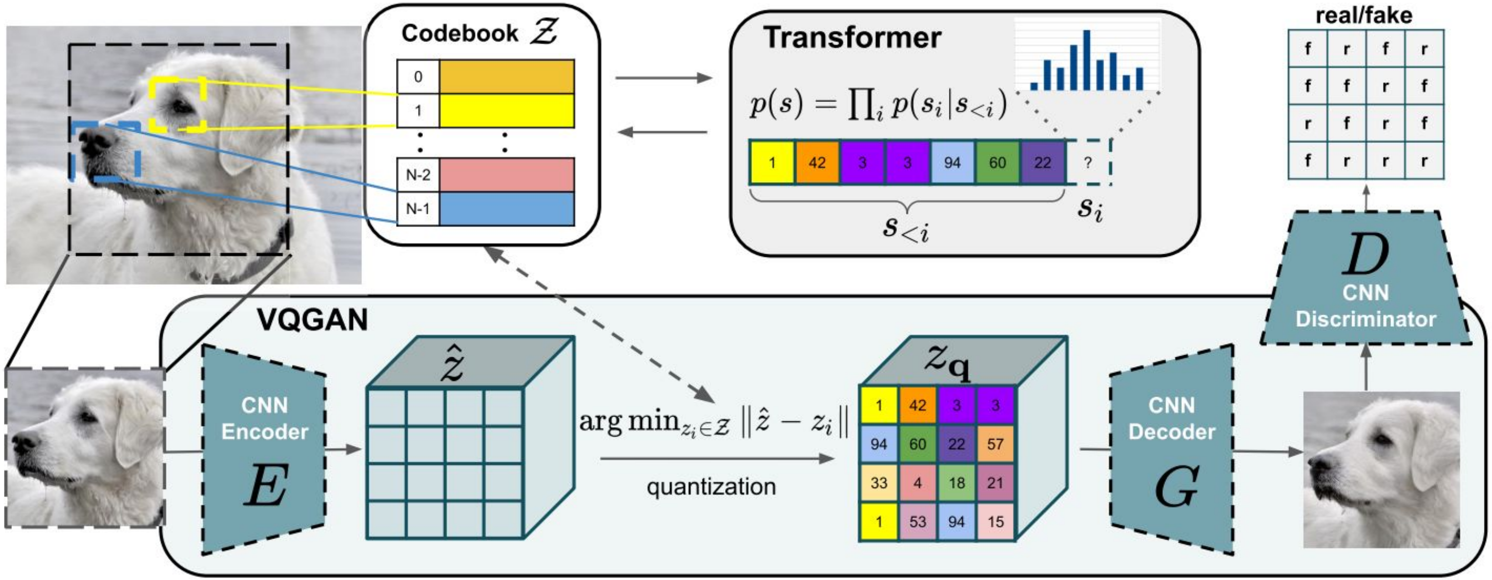


Figure 3: Samples (128x128) from a VQ-VAE with a PixelCNN prior trained on ImageNet images. From left to right: kit fox, gray whale, brown bear, admiral (butterfly), coral reef, alp, microwave, pickup.

Oord et al. Neural Discrete Representation Learning, 2017

VQ-GAN



Esser, Rombach et al. Taming Transformers for High-Resolution Image Synthesis, 2020

VQ-GAN

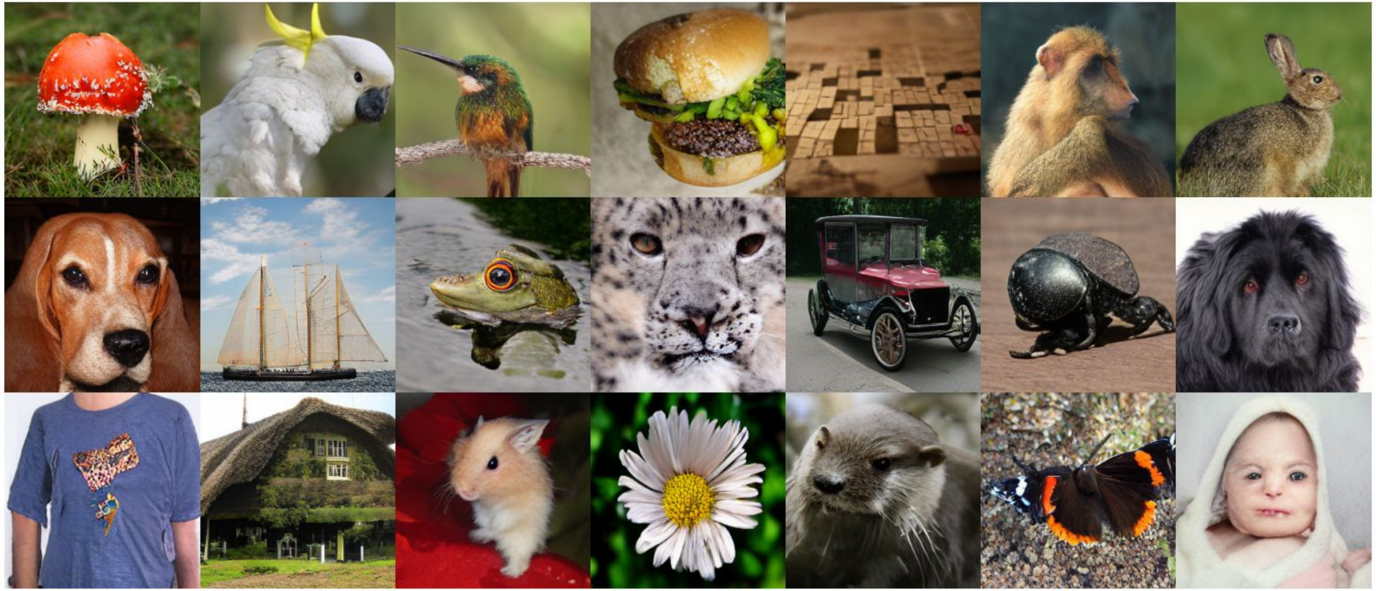


Figure 8. Samples from our class-conditional ImageNet model trained on 256×256 images.

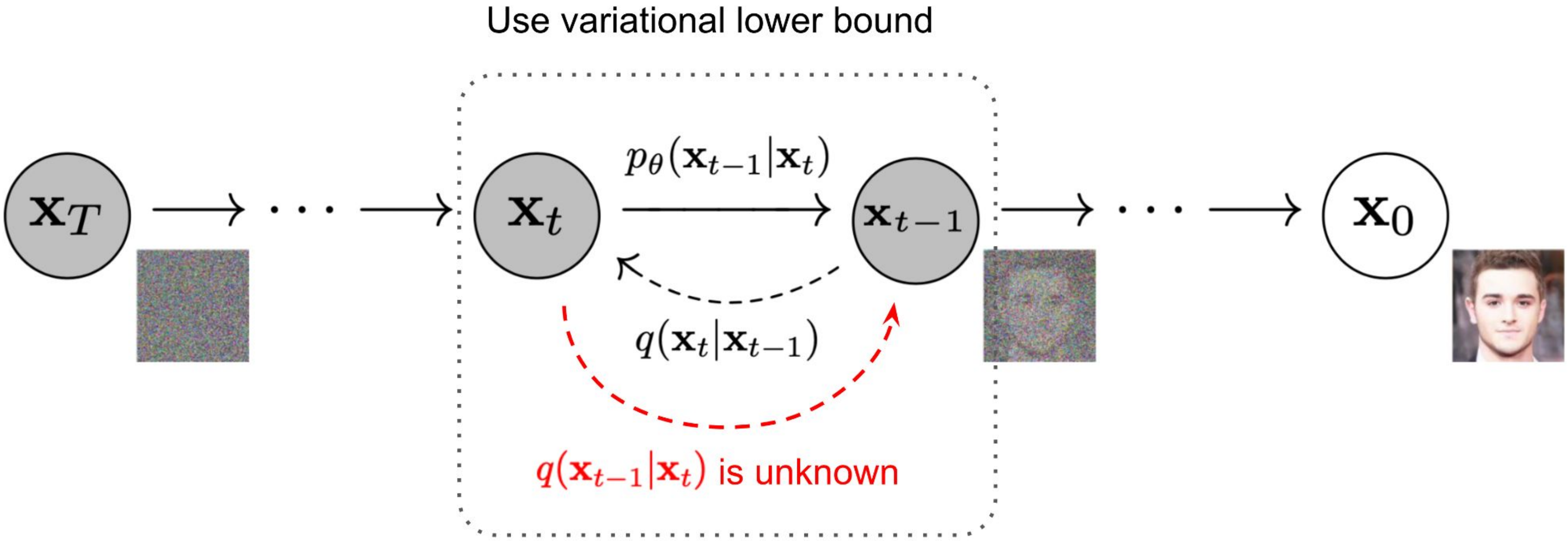
Esser, Rombach et al. Taming Transformers for High-Resolution Image Synthesis, 2020

VQ-GAN

Figure 1. Our approach enables transformers to synthesize high-resolution images like this one, which contains 1280x460 pixels.

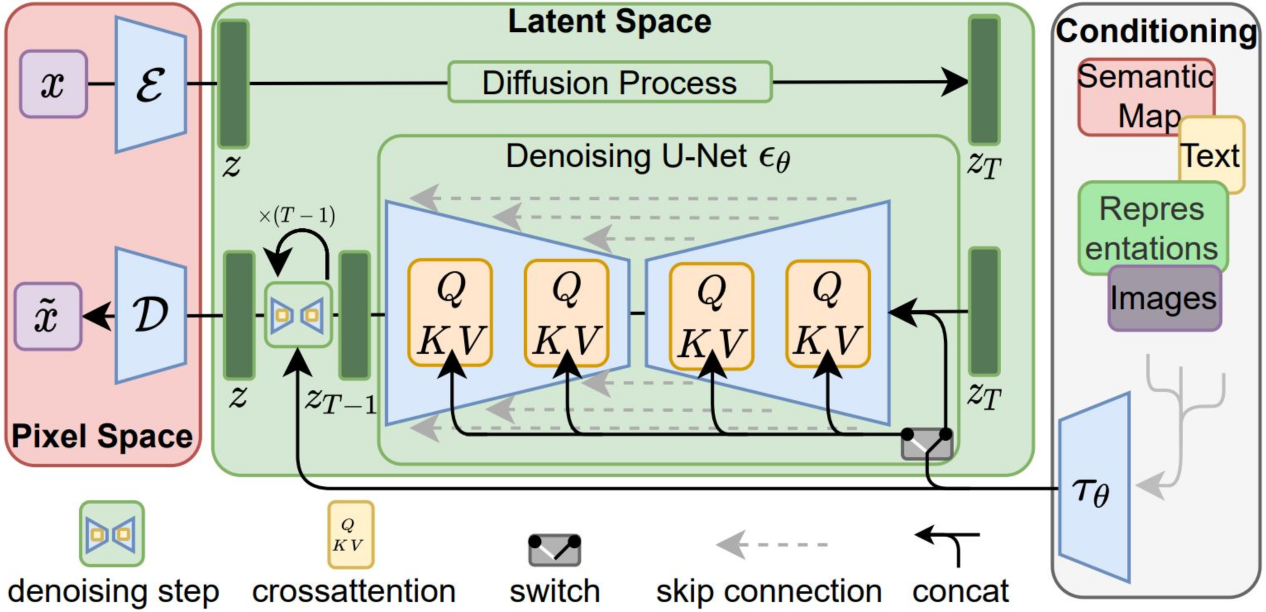
Esser, Rombach et al. Taming Transformers for High-Resolution Image Synthesis, 2020

Recap: diffusion models



Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics, 2015;
Yang & Ermon, 2019; DDPM; Ho et al. 2020; ... ?

Latent diffussions



Rombach, Blattmann et al. High-Resolution Image Synthesis with Latent Diffusion Models, 2021

Latent diffussions

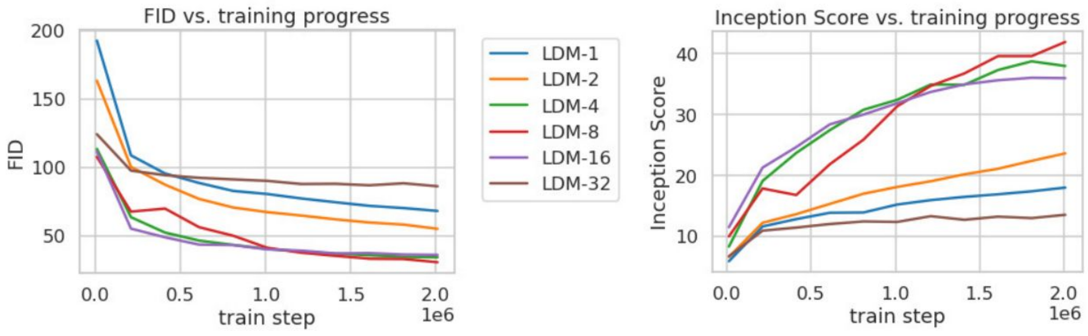
Training phases:

- Autoencoder
  - Loss: Patch-based GAN loss + Perceptual loss
  - Regularization: KL-loss (close to VAE) OR quantization in Decoder (like VQ-GAN)

- Various generative tasks
  - Loss: classical diffusion L2 restoration loss
  - All trainings done on single A100

Rombach, Blattmann et al. High-Resolution Image Synthesis with Latent Diffusion Models, 2021

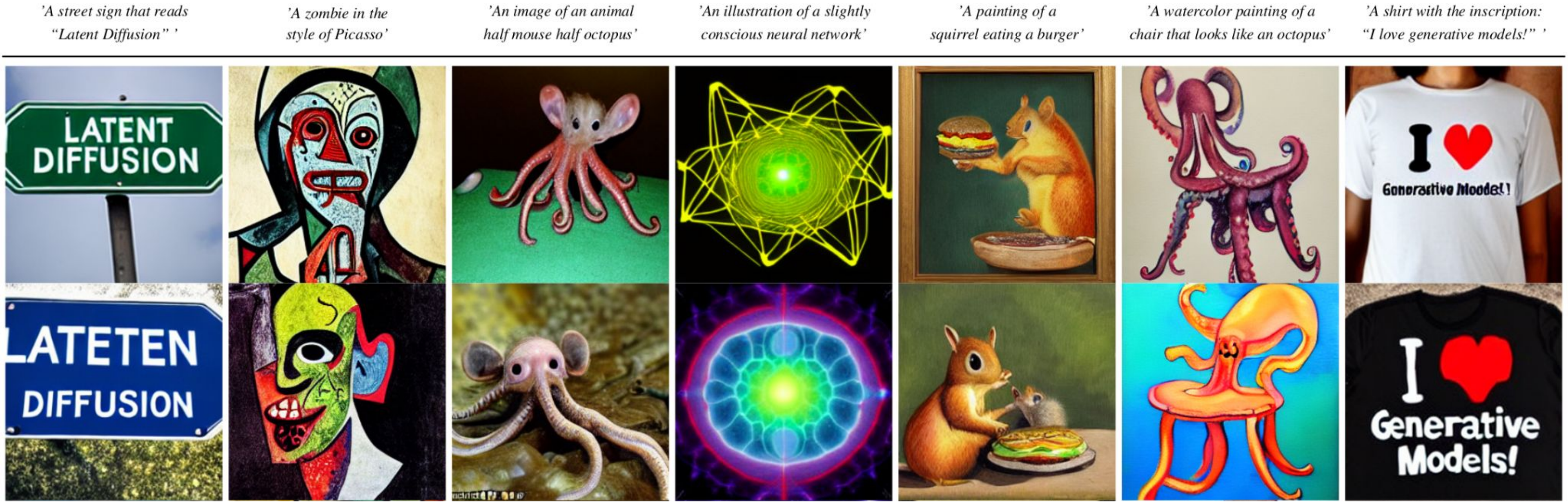
Latent diffussions



Downsampling for 4-16x: speedup of generative training without sampling quality loss

Rombach, Blattmann et al. High-Resolution Image Synthesis with Latent Diffusion Models, 2021

Latent diffussions



Rombach, Blattmann et al. High-Resolution Image Synthesis with Latent Diffusion Models, 2021

Latent diffussions

Figure 8. A *LDM* trained on 256^2 resolution can generalize to larger resolution (here: 512×1024) for spatially conditioned tasks such as semantic synthesis of landscape images. See Sec. 4.3.2.

Latent diffussions

Latent diffussions

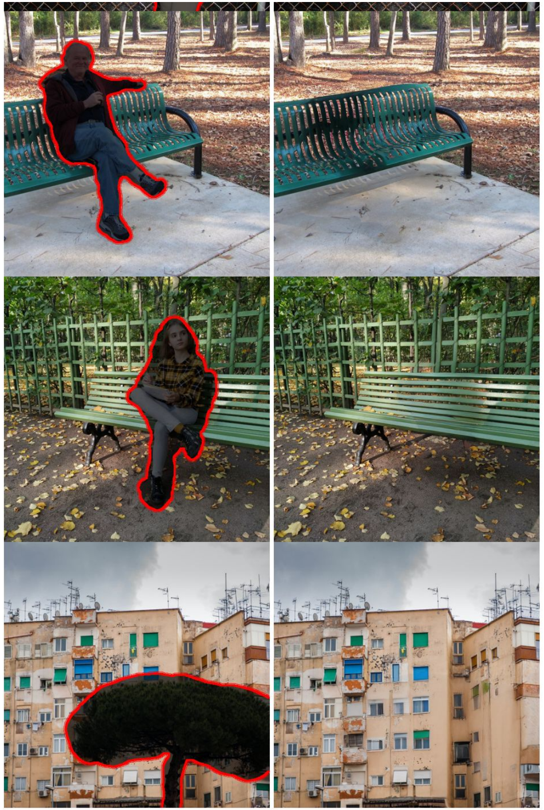


Figure 12. Qualitative results on object removal with our *big, w/ ft* inpainting model. For more results, see Fig. 22.

Latent diffussions

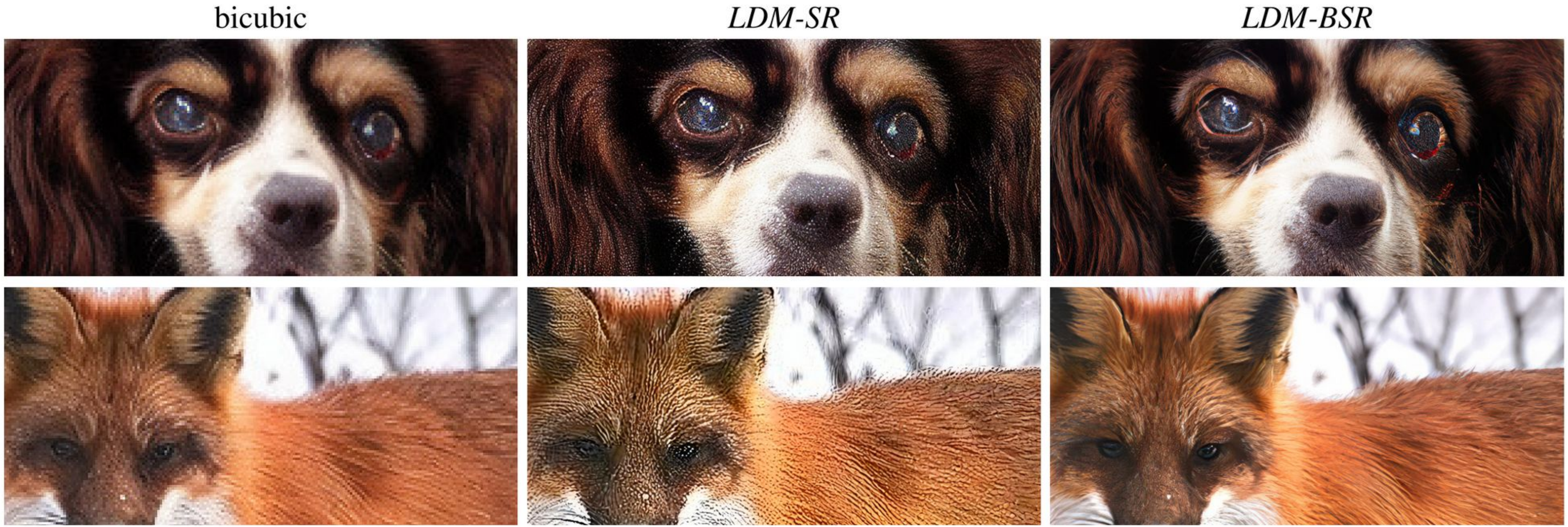
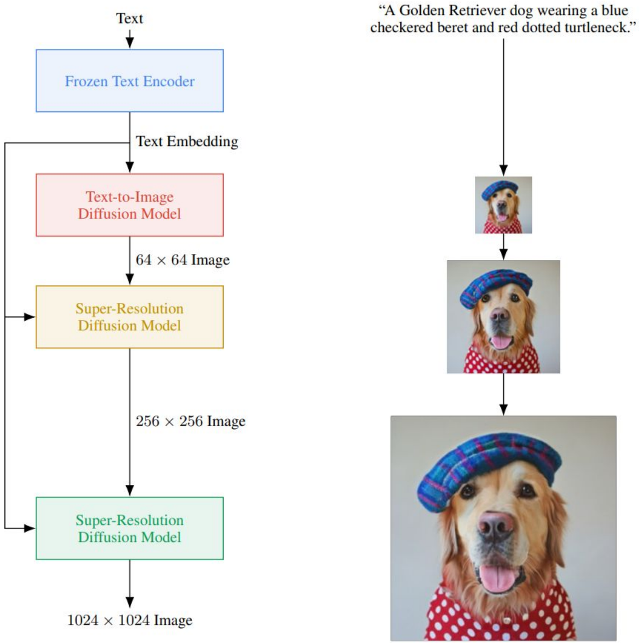


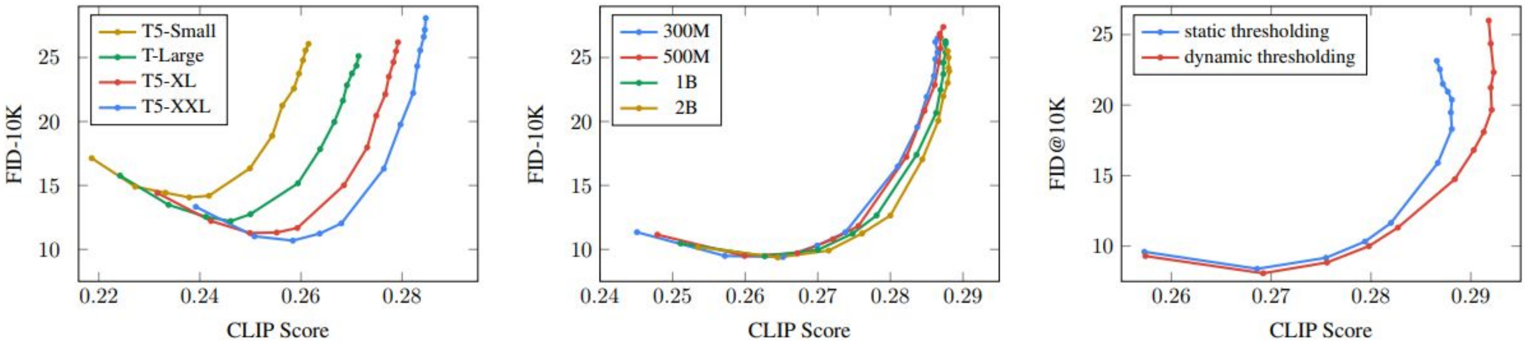
Figure 9. *LDM-BSR* generalizes to arbitrary inputs and can be used as a general-purpose upsampler, upscaling samples from a class-conditional *LDM* (image *cf*. Fig. 4) to 1024^2 resolution. In contrast, using a fixed degradation process (see Sec. 4.4) hinders generalization.

Imagen (important influencer)



Saharia, Chan. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, 2022

Imagen (important influencer)



(a) Impact of encoder size. (b) Impact of U-Net size. (c) Impact of thresholding.

Figure 4: Summary of some of the critical findings of Imagen with pareto curves sweeping over different guidance values. See Appendix D for more details.

Saharia, Chan. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, 2022

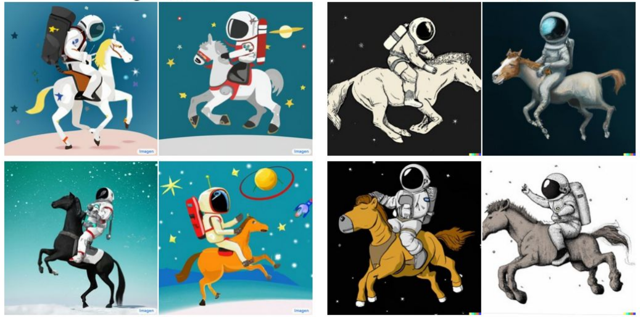
Imagen (important influencer)

Imagen (Ours) DALL-E 2 [54]

A yellow book and a red vase.

A black apple and a green backpack.

Imagen (Ours) DALL-E 2 [54]



A horse riding an astronaut.

A panda making latte art.

Figure A.19: Example qualitative comparisons between Imagen and DALL-E 2 [54] on DrawBench prompts from Conflicting category. We observe that both DALL-E 2 and Imagen struggle generating well aligned images for this category. However, Imagen often generates some well aligned samples, e.g. “A panda making latte art.”.

Stable diffusion

- No paper (core team is from Latent diffusion authors)
- Open source: https://github.com/CompVis/stable-diffusion
- Well improved and well trained latent diffusion model

Stable diffusion

Key components:

- High quality decoder from VQ-GAN
- Diffusion in latent space
- Frozen language model (CLIP ViT-L/14 embeddings)
- Classifier-free guidance
- A lot of data (LAION-5B and its subsets)
- A lot of compute power

Stable diffusion