High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion)

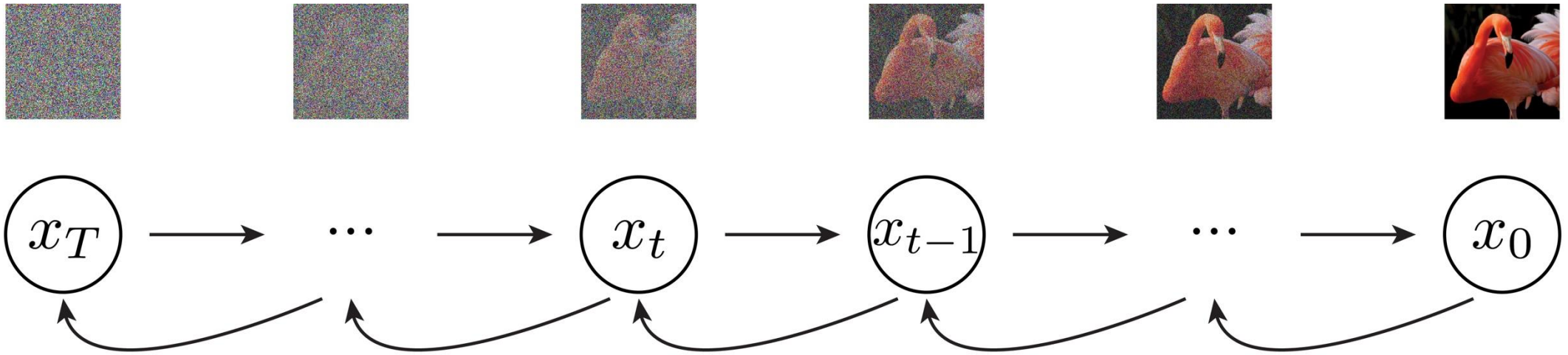
Robin Rombach, Andreas Blattmann, Dominik Lorenz,
Patrick Esser, and Björn Ommer

CVPR 2022



A chair that looks like an octopus

Diffusion in one minute



LDM keeps this learning problem and moves it into latent space.

Why diffuse in a latent space?

Pixel diffusion repeats expensive computation on a large image grid.

Pixel grid

512 × 512

262,144 spatial positions

Latent grid

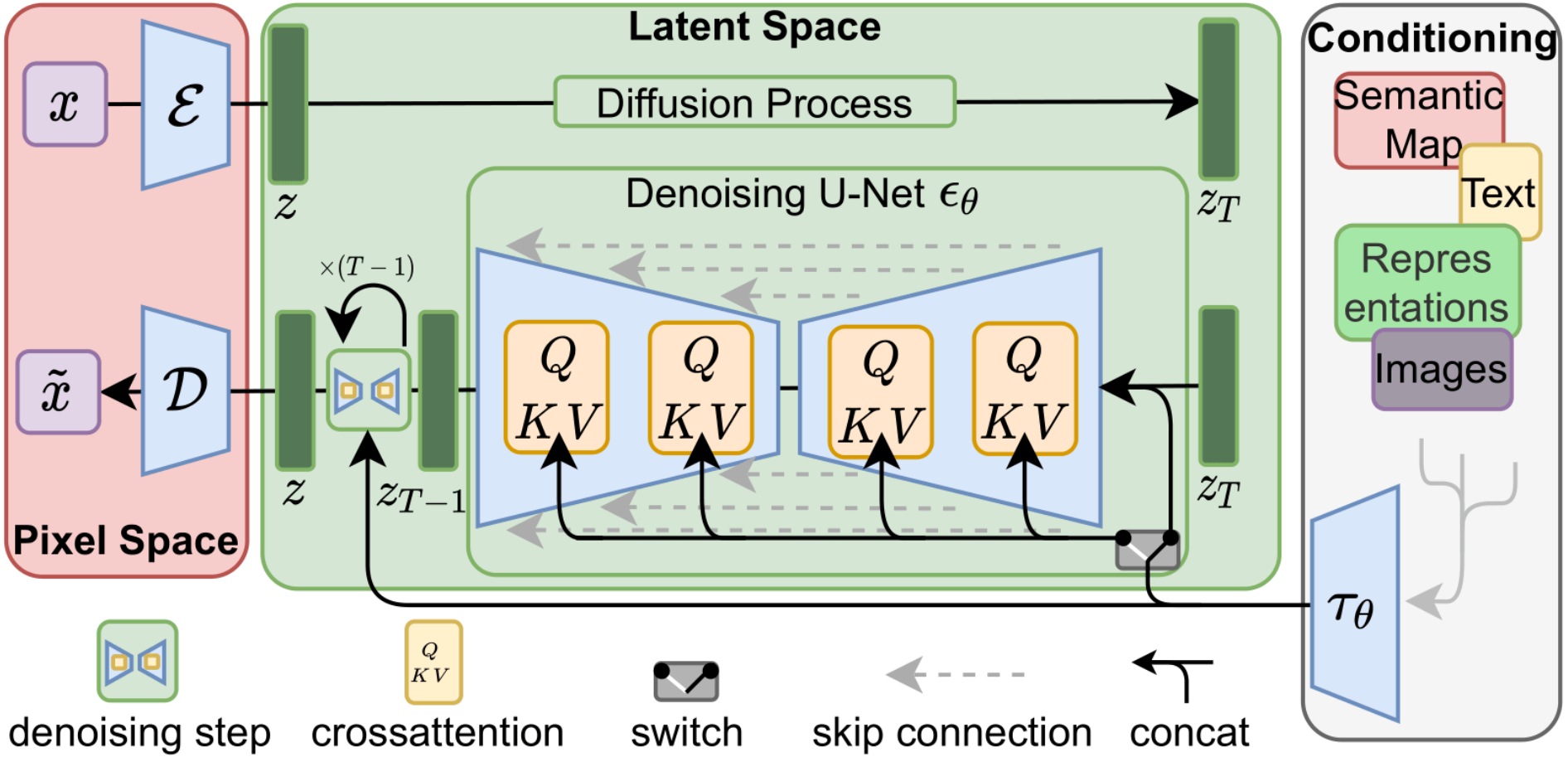
64 × 64

4,096 spatial positions

64× fewer spatial positions

The actual speedup also depends on channels, architecture, and sampling.

The two-stage model



1 Learn the autoencoder

2 Freeze it and train latent diffusion

Stage 1: perceptual compression

$$z = \mathcal{E}(x) \qquad \tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$$



- Preserve visible structure and texture in a compact spatial tensor
- VAE (continuous) or VQGAN (discrete)

Stage 2: learn to denoise latent codes

Forward noising

$$q(x_t|x_0) = \mathcal{N}(x_t|\alpha_t x_0, \sigma_t^2 \mathbb{I})$$

The paper uses x here.
For LDMs, apply this to latent z.

Conditional training objective

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

Train the conditioned U-Net to predict the added noise.

Update the U-Net and condition encoder. Keep E and D fixed.

Text conditioning with cross-attention

"A painting of a squirrel eating a burger"

Queries Q

Spatial features
from the U-Net

Keys K and values V

Token features
from the text encoder

$\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$, which is then mapped to the intermediate layers of the UNet via a cross-attention layer implementing $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$, with

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y).$$

Sampling from a text prompt

1 **Encode the prompt**

2 **Denoise a random latent**

Repeat the U-Net with the prompt.

3 **Decode once to pixels**

“A painting of a squirrel eating a burger”

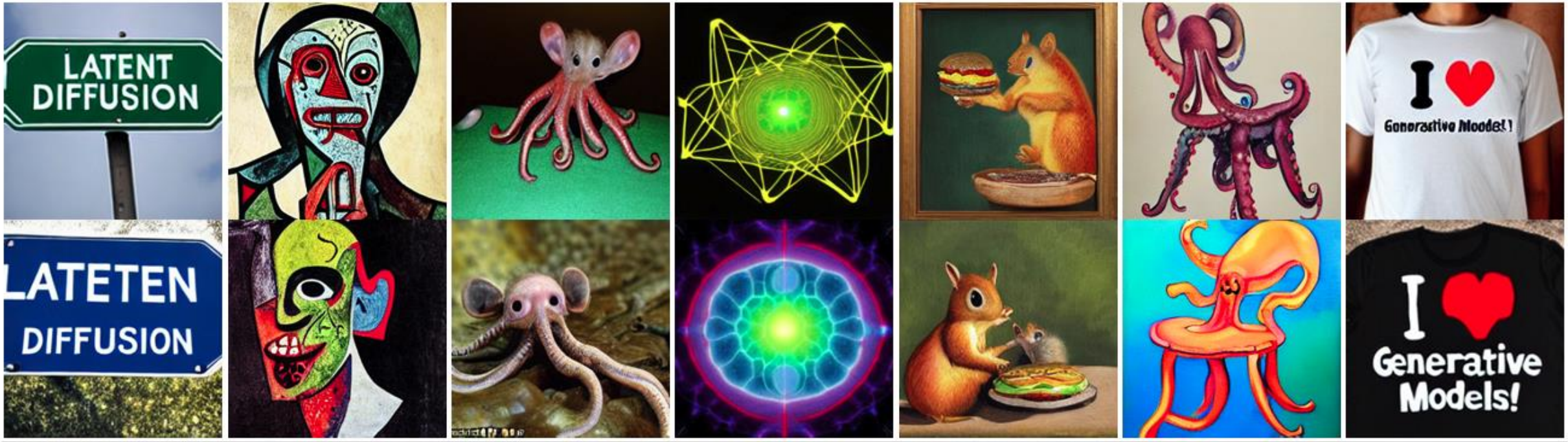


The random seed changes the result.

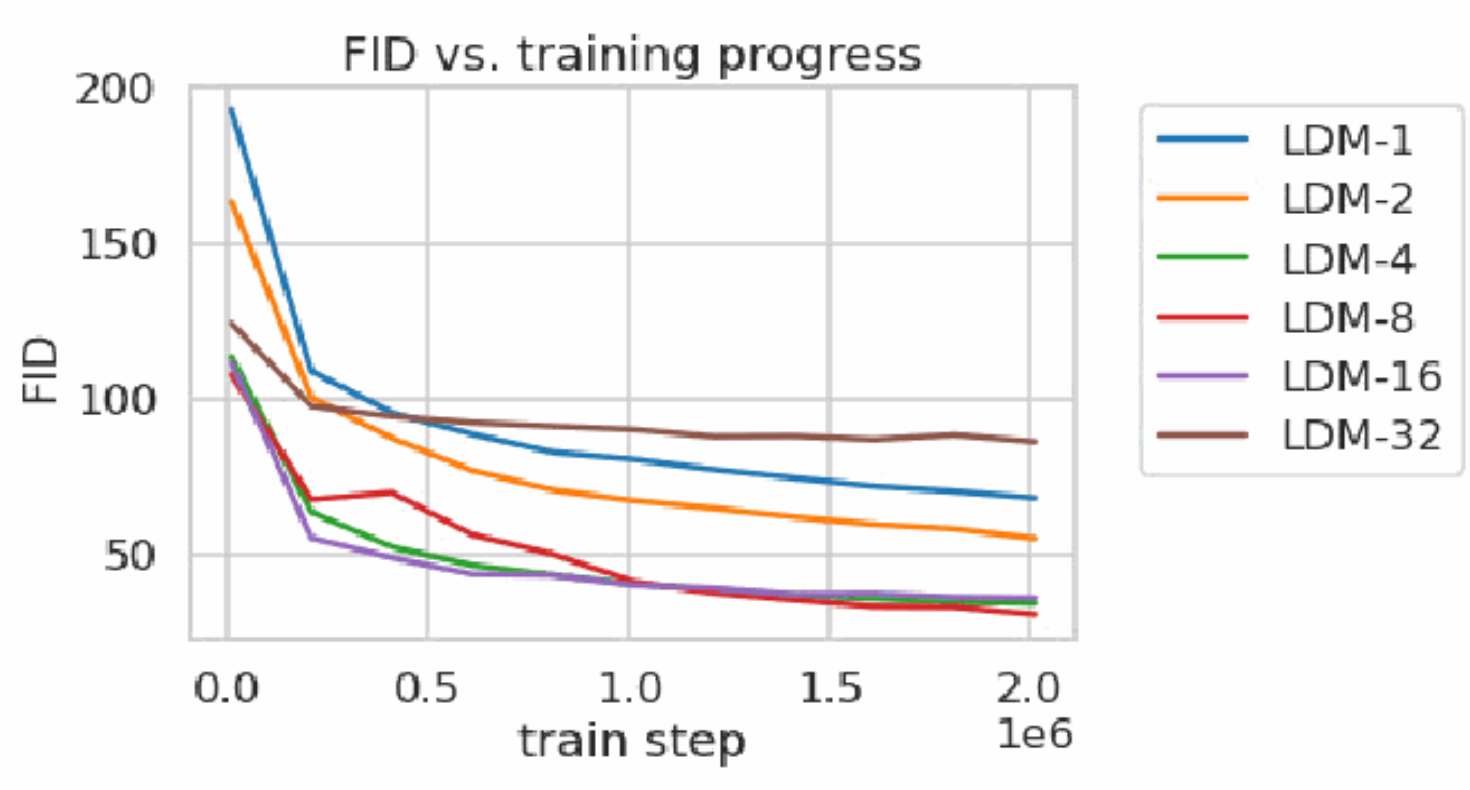
Results

Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads "Latent Diffusion" '	*'A zombie in the style of Picasso'*	*'An image of an animal half mouse half octopus'*	*'An illustration of a slightly conscious neural network'*	*'A painting of a squirrel eating a burger'*	*'A watercolor painting of a chair that looks like an octopus'*	*'A shirt with the inscription: "I love generative models!" '*



Compression controls the quality–cost tradeoff



FID ↓ is better

Too little compression

Large grids make each step expensive.

Too much compression

Lost information limits sample quality.

f = 4 or 8 works well in these experiments.

Unconditional image synthesis benchmark

CelebA-HQ 256 × 256				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T. [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49	*LDM-4* (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN [41]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.59**	0.61	0.44	ProjectedGAN [76]	**1.52**	0.61	0.34
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**	*LDM-4* (ours, 200-s)	2.95	**0.66**	0.48

Table 1. Evaluation metrics for unconditional image synthesis.

text-to-image benchmark

Text-Conditional Image Synthesis				
Method	FID ↓	IS↑	N_{params}	
CogView† [17]	27.10	18.20	4B	self-ranking, rejection rate 0.017
LAFITE† [109]	26.94	<u>26.02</u>	75M	
GLIDE* [59]	<u>12.24</u>	-	6B	277 DDIM steps, c.f.g. [32] $s = 3$
Make-A-Scene* [26]	**11.84**	-	4B	c.f.g for AR models [98] $s = 5$
LDM-KL-8	23.31	$20.03_{\pm 0.33}$	1.45B	250 DDIM steps
*LDM-KL-8-G**	12.63	$\mathbf{30.29}_{\pm \mathbf{0.42}}$	1.45B	250 DDIM steps, c.f.g. [32] $s = 1.5$

Table 2. Evaluation of text-conditional image synthesis on the 256×256-sized MS-COCO [51] dataset: with 250 DDIM [84]

A measured gain in efficiency

Inpainting: pixel diffusion versus latent diffusion

Model	Train samples / sec	FID ↓ after 6 epochs
Pixel diffusion (f = 1)	0.11	24.74
Latent diffusion (f = 4)	0.33	14.99

3× training throughput

Better sample quality in this reported comparison.

Layout-to-image synthesis



High resolution through spatial conditioning



Training images

256 × 256

Generated image

512 × 1024

Semantic map in the inset supplies the scene layout.

Inpainting



Method	40-50% masked		All samples	
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
LDM-4 (ours, big, w/ ft)	**9.39**	0.246± 0.042	**1.50**	0.137± 0.080
LDM-4 (ours, big, w/o ft)	12.89	0.257± 0.047	2.40	0.142± 0.085
LDM-4 (ours, w/ attn)	11.87	0.257± 0.042	2.15	0.144± 0.084
LDM-4 (ours, w/o attn)	12.60	0.259± 0.041	2.37	0.145± 0.084
LaMa [88]†	12.31	**0.243**± 0.038	2.23	**0.134**± 0.080
LaMa [88]	12.0	**0.24**	2.21	0.14
CoModGAN [107]	10.4	0.26	1.82	0.15
RegionWise [52]	21.3	0.27	4.75	0.15
DeepFill v2 [104]	22.1	0.28	5.20	0.16
EdgeConnect [58]	30.5	0.28	8.37	0.16

Table 7. Comparison of inpainting performance on 30k crops of

Questions

- Good reconstruction does not always lead to good generation. What are some useful ways to understand this gap? Are there any promising approaches to design a latent space that works well for both reconstruction and generation?
- How come LDM outperforms other diffusion models? What differences were there that made this method better, considering that the downsampling with the autoencoder is unlikely to be the reason for the better performance?
- Could the initial perceptual compression step inadvertently throw away information