Latent Diffusion Model

(stable diffusion)

Artem Kushneruk, 04.03.2024

Problem

- Training diffusion models consumes hundreds of GPU days
- Inference is expensive due to sequential evaluations (e.g 50k samples take 5 days on a single A100 GPU

Goal: reduce computational complexity

Ideas

Run **diffusion** in the **latent space** using autoencoder (pretrained on our data)**:**

- Compress with encoder
- Run diffusion process
- Decompress with decoder

Result: near-optimal point between complexity reduction and details preservation

Autoencoder

$x \in \mathscr{R}^{H \times W \times 3}, z \in \mathscr{R}^{h \times w \times c}, , z = \mathscr{E}(x)$

$f = H/h = W/w = 2^m$ – downsampling factor

Important: we have 2D representation, so image-specific inductive bias can be applied (U-net)

Avoiding high-variance in the latent space

[training autoencoder]

- KL-reg: A small KL penalty towards a $\mathcal{N}(0,1)$ over the learned latent (like in VAE)
- VQ-reg: Vector quantisation layer within the decoder (like in VQGAN)

Diffusion models

Diffusion models learn data distribution p(x) by denoising a $\mathcal{N}(?, ?)$ by learning reverse process of a fixed Markov Chain

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big]\,,$$

with t uniformly sampled from $\{1, \ldots, T\}$.

Key moments

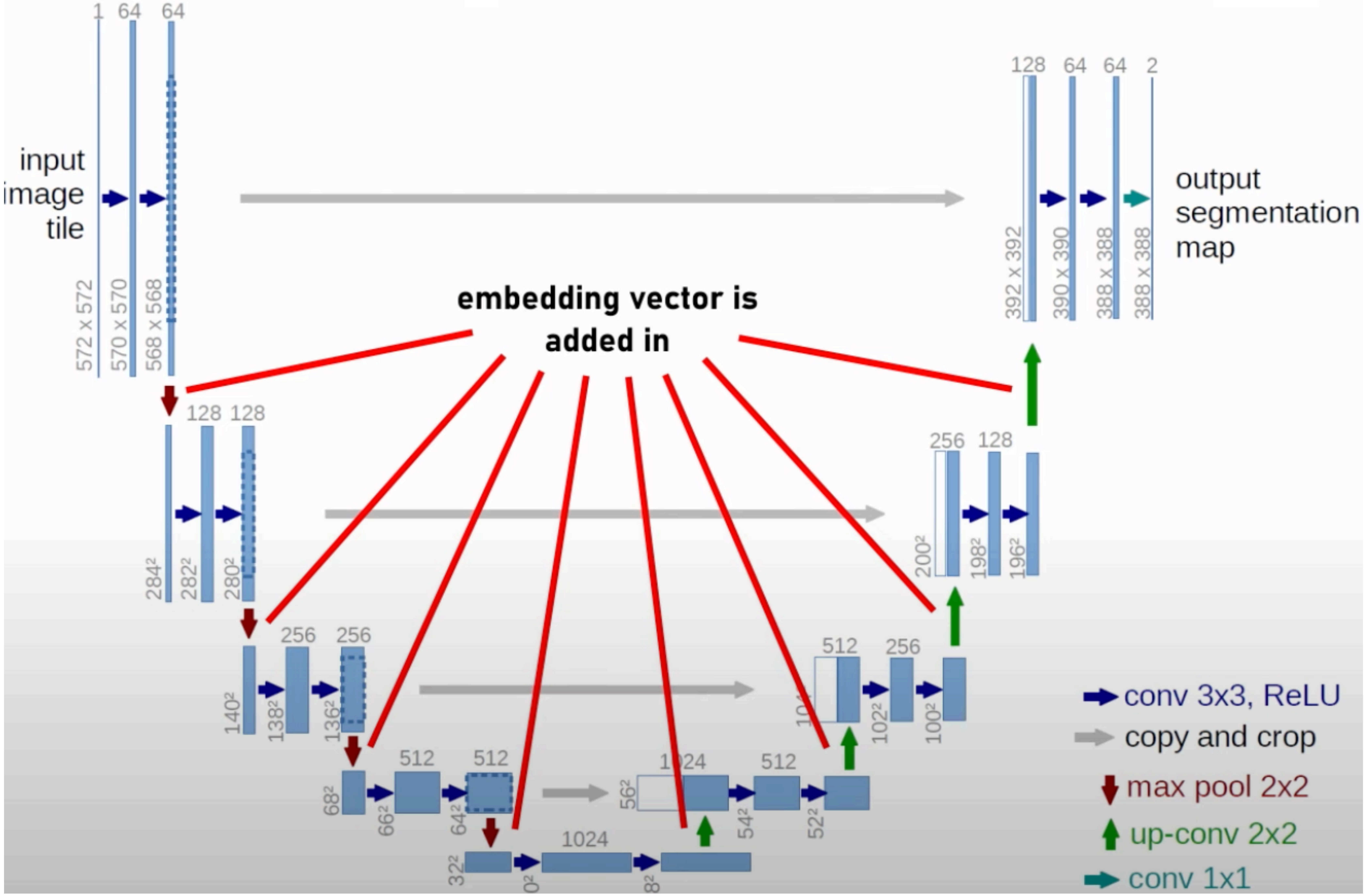
low-dimensional latent space allows:

- model can focus on important bits of data (no high frequency details in data)
- model can train in computationally efficient space

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

$\epsilon_\theta(\, . \, , t)$ — time conditional UNet

Time conditional UNet



Conditioning

DMs can model $p(z \mid y)$ by modeling $\epsilon_\theta(z_t, t, y)$ with NNs

Authors use UNet with cross-attention mechanism.

$\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$ - cond. inf. encoding (e.g CLIP)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \text{ with}$$

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

$\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ — flattened representation from UNet

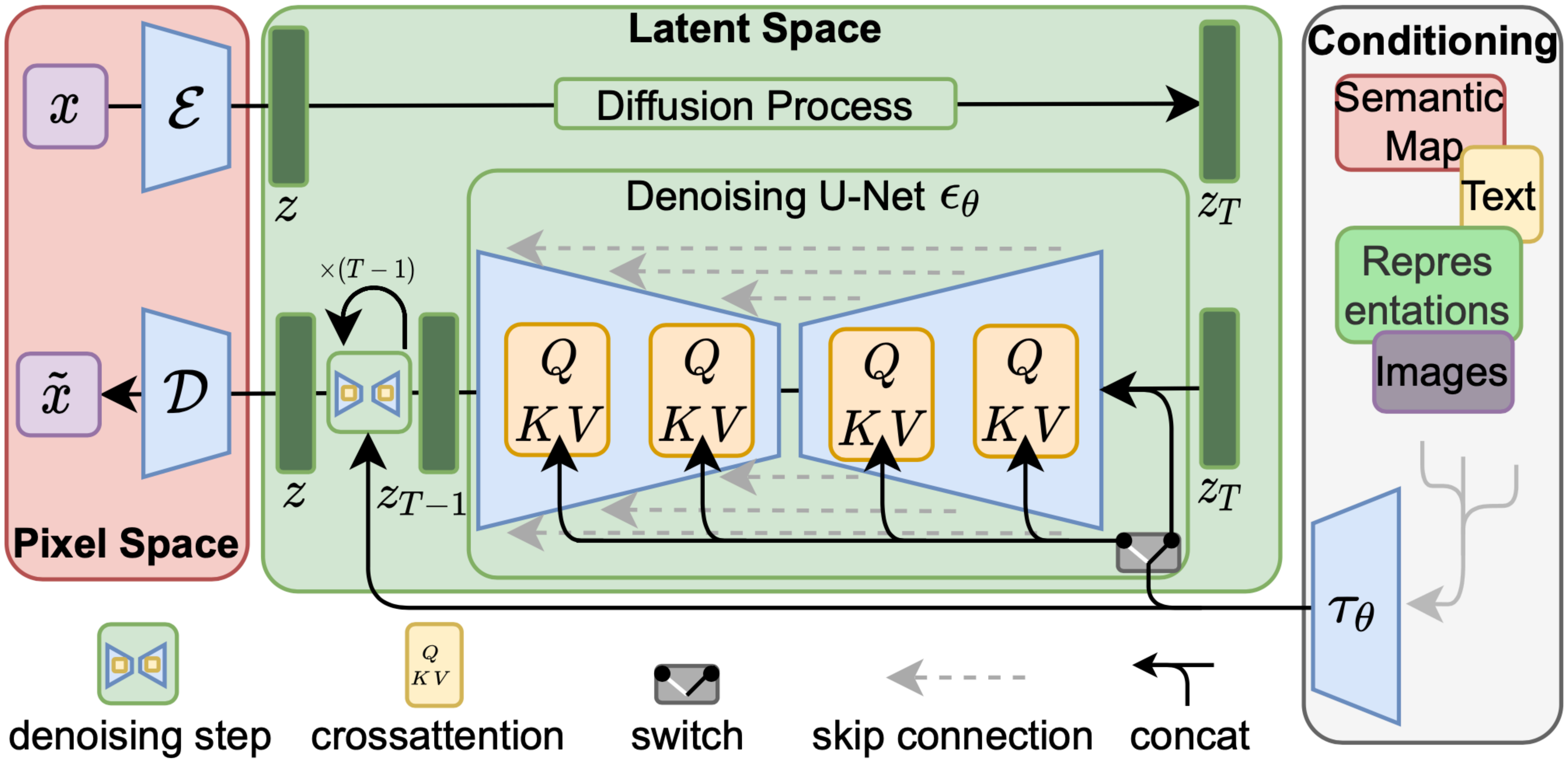
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\Big]$$

Denoising model

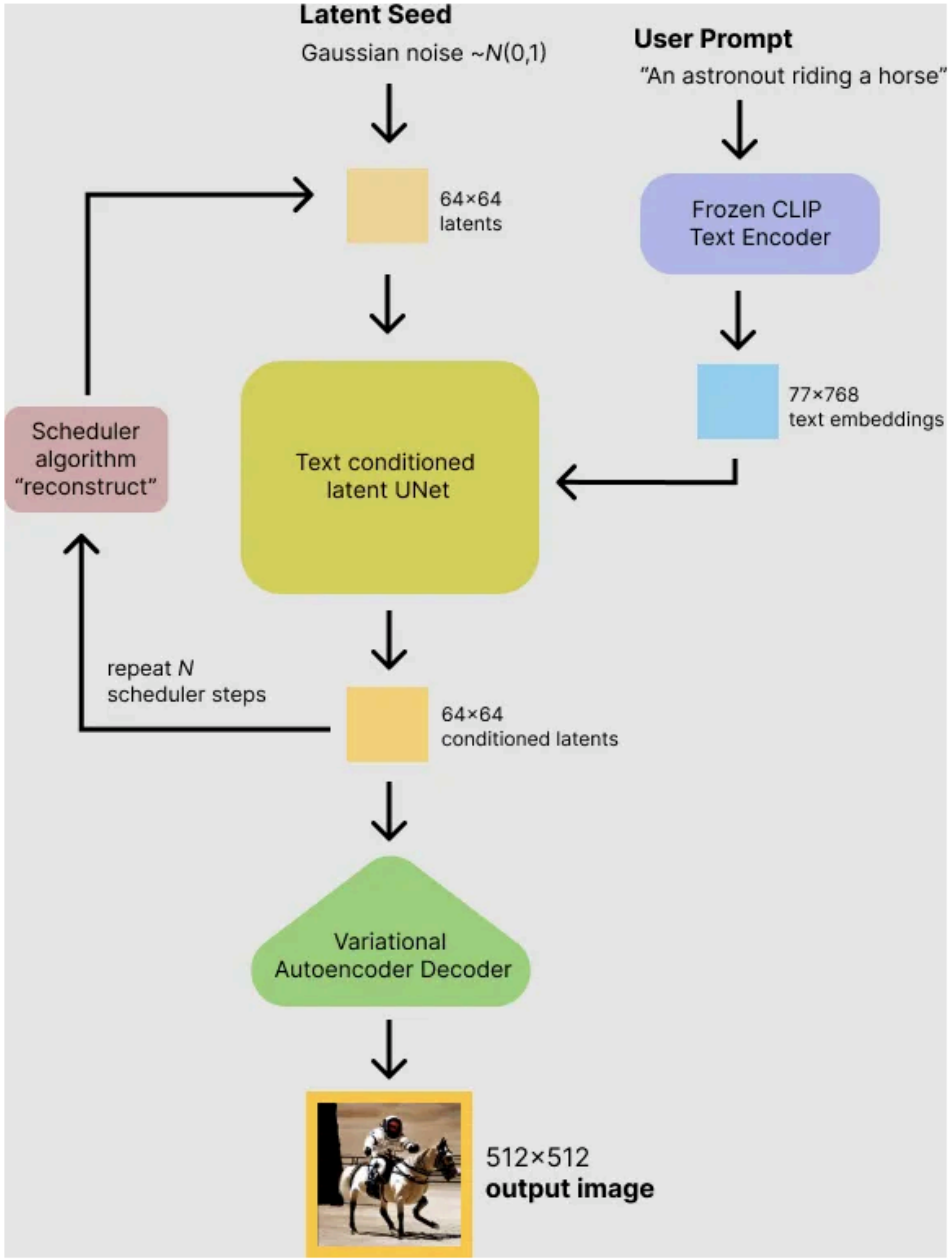
U-net with cross-attention mechanism for conditional information (CI)

For *CI* use domain-specific encoder

Attention: queries are latent pictures, keys and values from encoded *CI*



Putting it all together, the model works as follow during inference process:



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!" '*

Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

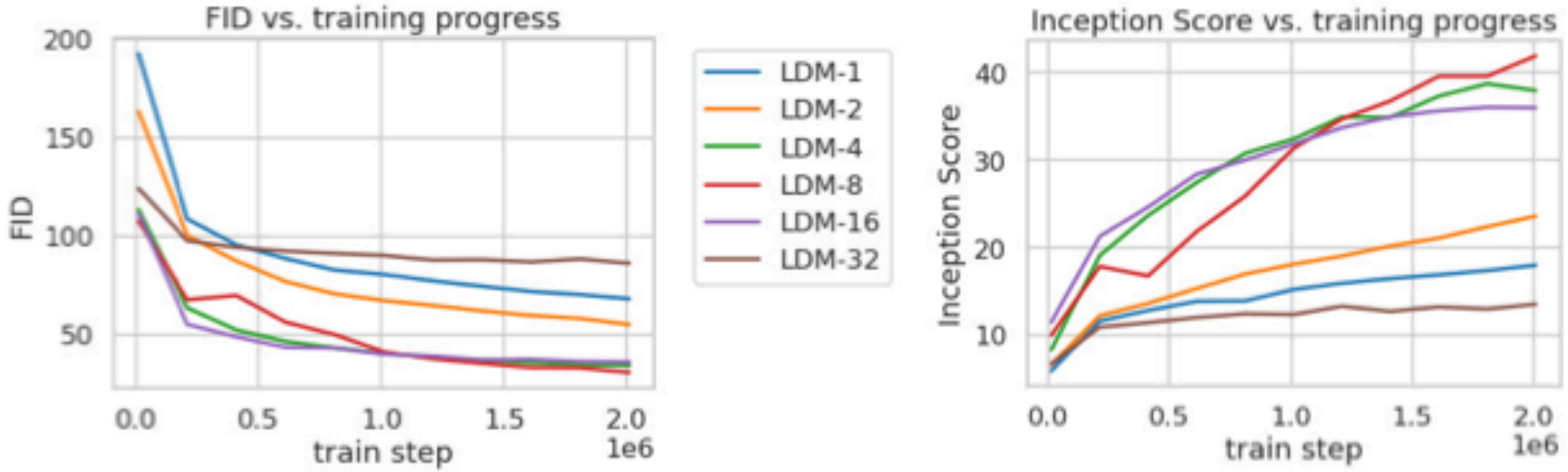


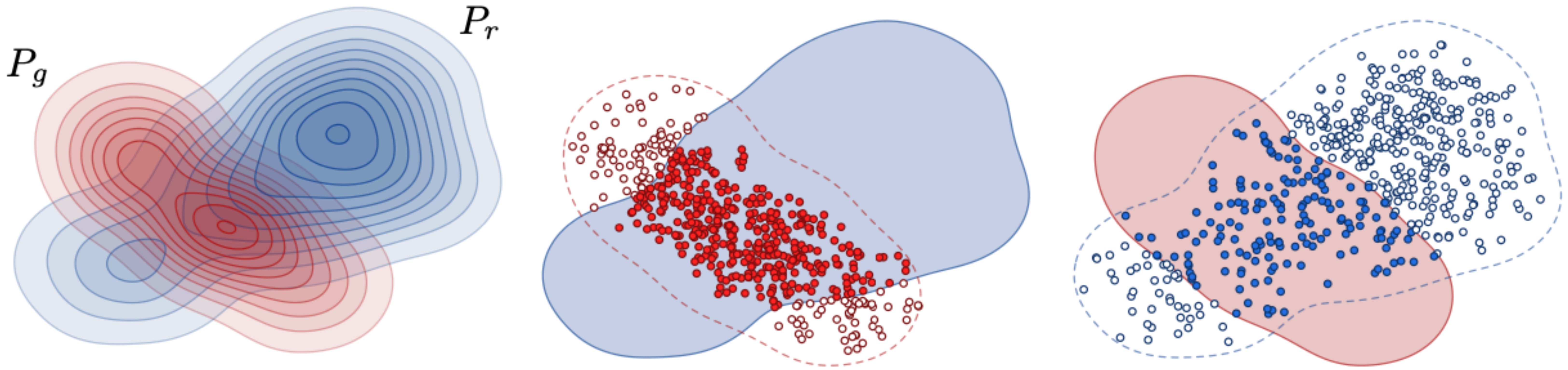
Figure 6. Analyzing the training of class-conditional *LDMs* with different downsampling factors f over 2M train steps on the ImageNet dataset. Pixel-based *LDM-1* requires substantially larger train times compared to models with larger downsampling factors (*LDM*-$\{4\text{-}16\}$). Too much perceptual compression as in *LDM-32* limits the overall sample quality. All models are trained on a single NVIDIA A100 with the same computational budget. Results obtained with 100 DDIM steps [84] and $\kappa = 0$.

CelebA-HQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-
VQGAN+T. [23] (k=400)	10.2	-	-
PGGAN [39]	8.0	-	-
LSGM [93]	7.22	-	-
UDM [43]	$\underline{7.16}$	-	-
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49

FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [21]	9.57	-	-
U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
UDM [43]	5.54	-	-
StyleGAN [41]	$\underline{4.16}$	$\underline{0.71}$	$\underline{0.46}$
ProjectedGAN [76]	**3.08**	0.65	$\underline{0.46}$
LDM-4 (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-
ImageBART [21]	7.32	-	-
PGGAN [39]	6.42	-	-
StyleGAN [41]	4.21	-	-
StyleGAN2 [42]	$\underline{3.86}$	-	-
ProjectedGAN [76]	**1.59**	$\underline{0.61}$	$\underline{0.44}$
LDM-8* (ours, 200-s)	4.02	**0.64**	**0.52**

LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [21]	5.51	-	-
DDPM [30]	4.9	-	-
UDM [43]	4.57	-	-
StyleGAN [41]	2.35	0.59	$\underline{0.48}$
ADM [15]	$\underline{1.90}$	**0.66**	**0.51**
ProjectedGAN [76]	**1.52**	$\underline{0.61}$	0.34
LDM-4 (ours, 200-s)	2.95	**0.66**	$\underline{0.48}$



(a) Example distributions (b) Precision (c) Recall

Classifier-free guidance

discard conditional data at random step, so $\epsilon_\theta(x_t, t) = \epsilon_\theta(x_t, t, y = \varnothing)$

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	$\mathbf{30.29}_{\pm 0.42}$	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] steps our model is on par with the most recent diffusion [59] and autoregressive [26] methods despite using significantly less parameters. †/*:Numbers from [109]/ [26]