Stable Diffusion

서울대학교 IDEA 연구실
석사 신윤섭

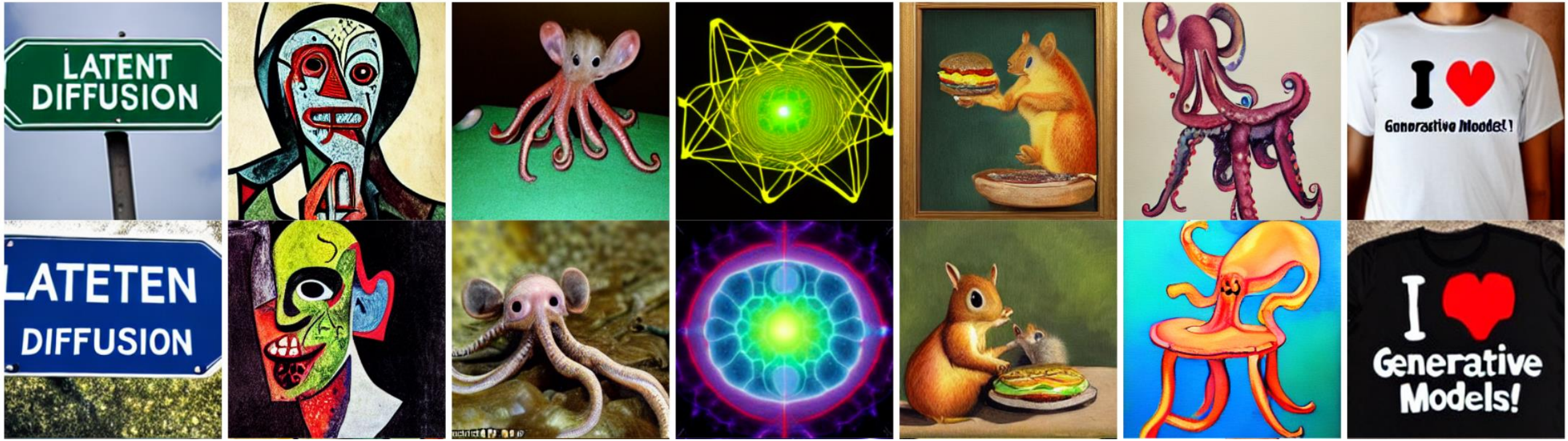
High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]

[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML

https://github.com/CompVis/latent-diffusion

'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!" '*



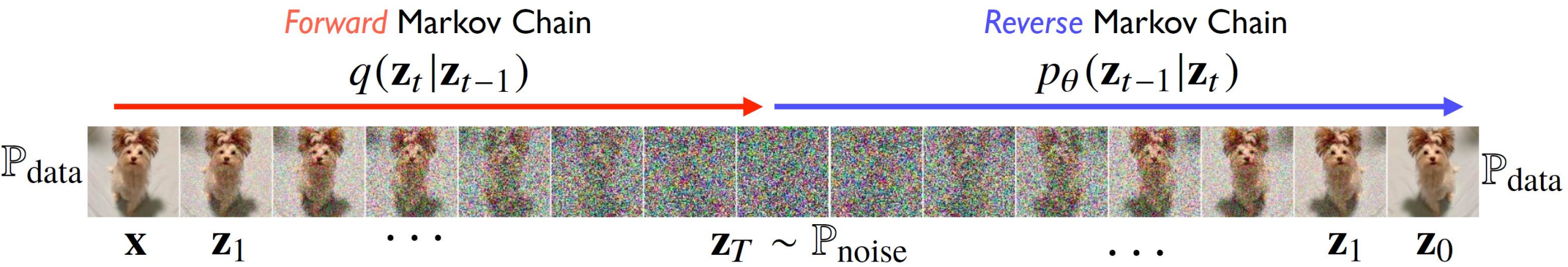
OUTLINE

01. Introduction

- The name 'Stable Diffusion' is widely known, deriving from the name of the company 'Stability AI' founded by the authors.
- This was published at CVPR 2022.
- This is a method proposed to compensate for the computational drawbacks of diffusion models.
- The code and pretrained model are released as open source, and they serve as the foundation for many image generation GUIs like Novel AI.

02. Prerequisite

- Before exploring LMD, let's briefly review diffusion models.



$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \qquad p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mu_\theta(t, \mathbf{z}_t), \Sigma_\theta(t, \mathbf{z}_t))$$

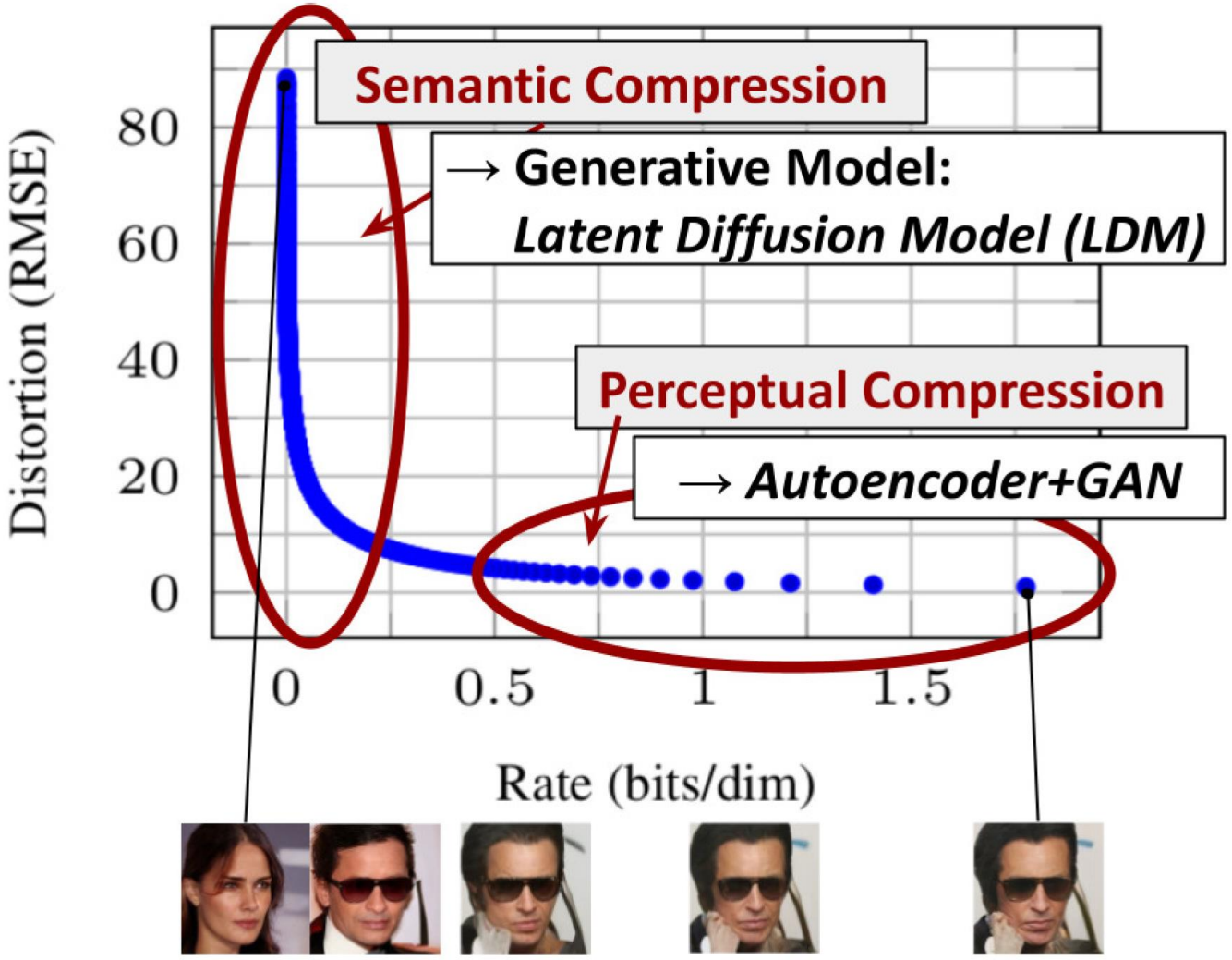
$$L_{DM} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$

Ho et al., Denoising Diffusion-Probabilistic Models, NeurIPS (2020)

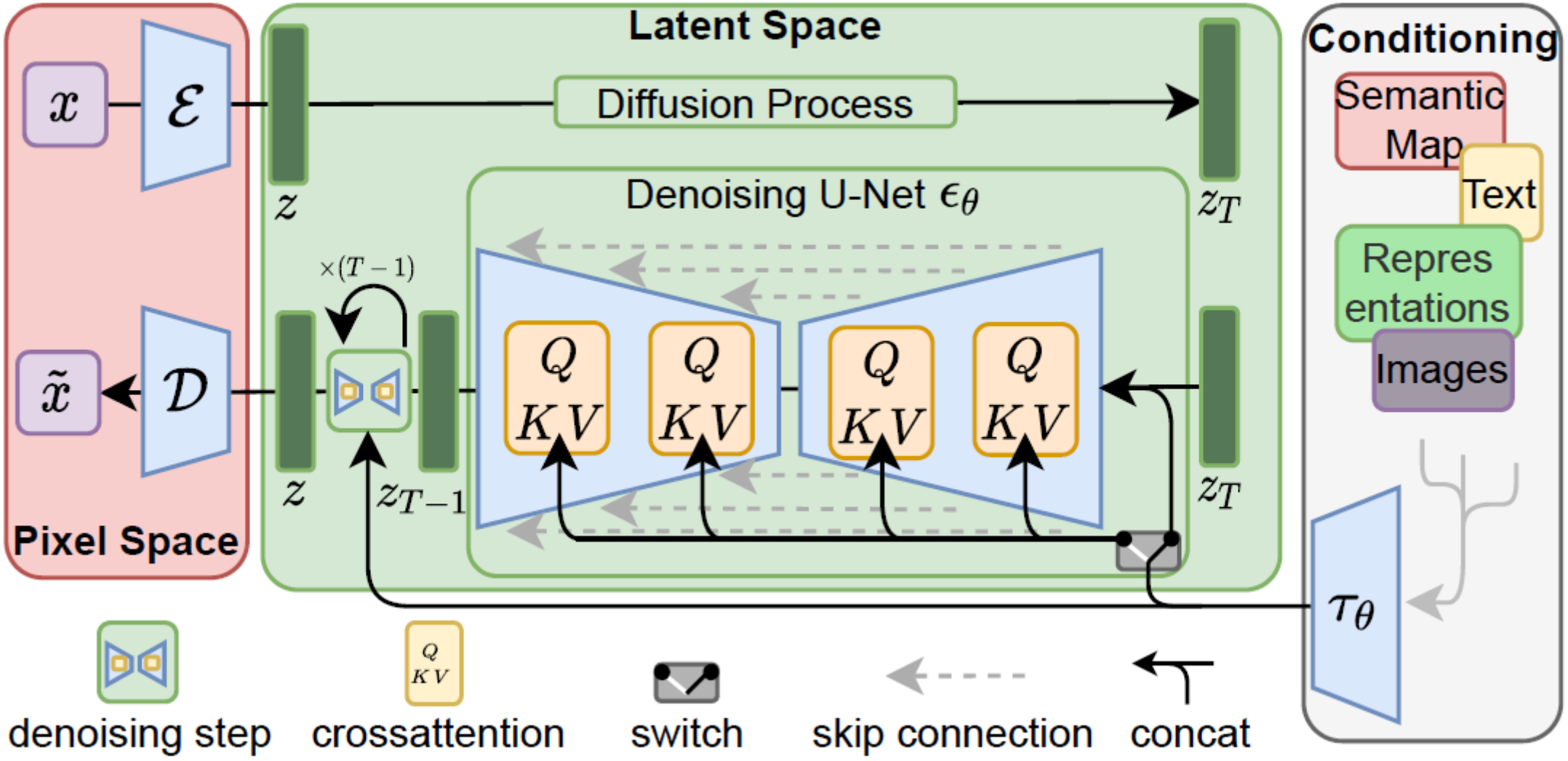
03. Latent Diffusion Model

- Since the diffusion model optimizes in RGB space, the dimensionality is high, and gradients must be calculated even for less important pixels, resulting in significant computational issues due to the large amount of computation required.
- To enable DM training on limited computational resources while retaining their quality and flexibility, LDM apply them in the latent space of powerful pretrained autoencoders.
- In other words, the original image is dimensionally reduced through an autoencoder, and then diffusion is applied in the reduced latent space.
- In this process, the autoencoder extracts the perceptual features of the image.

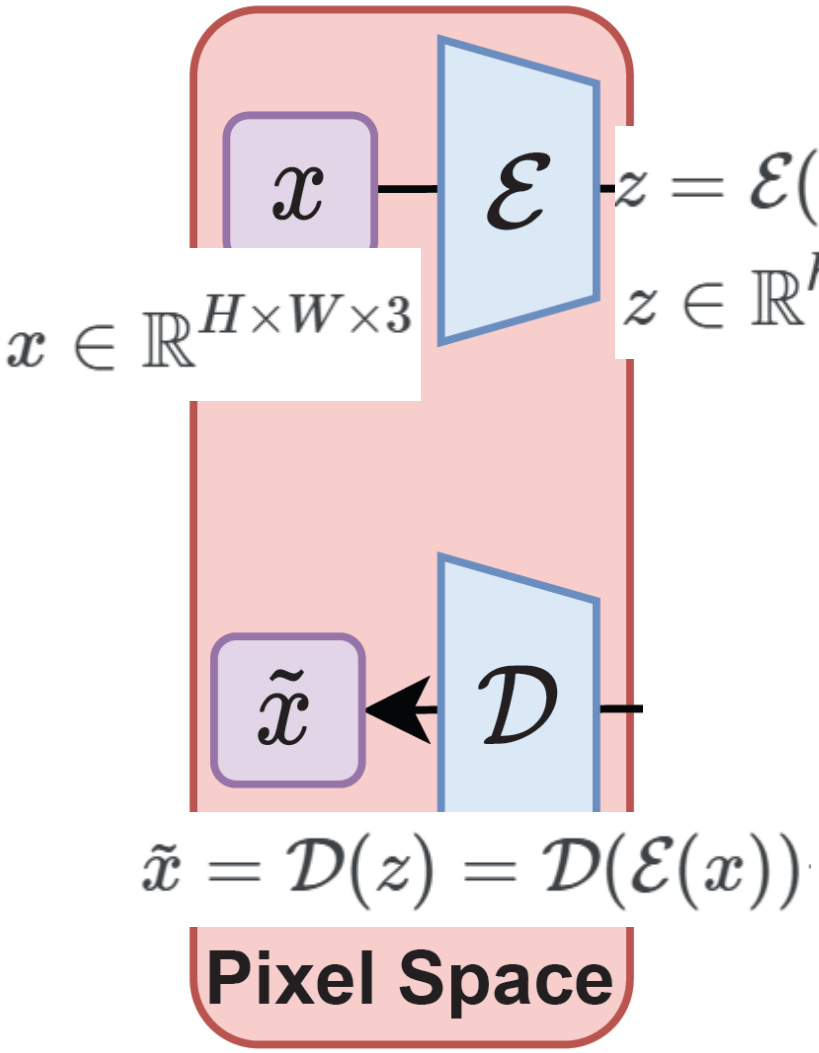
03. Latent Diffusion Model



03. Latent Diffusion Model



03. Latent Diffusion Model

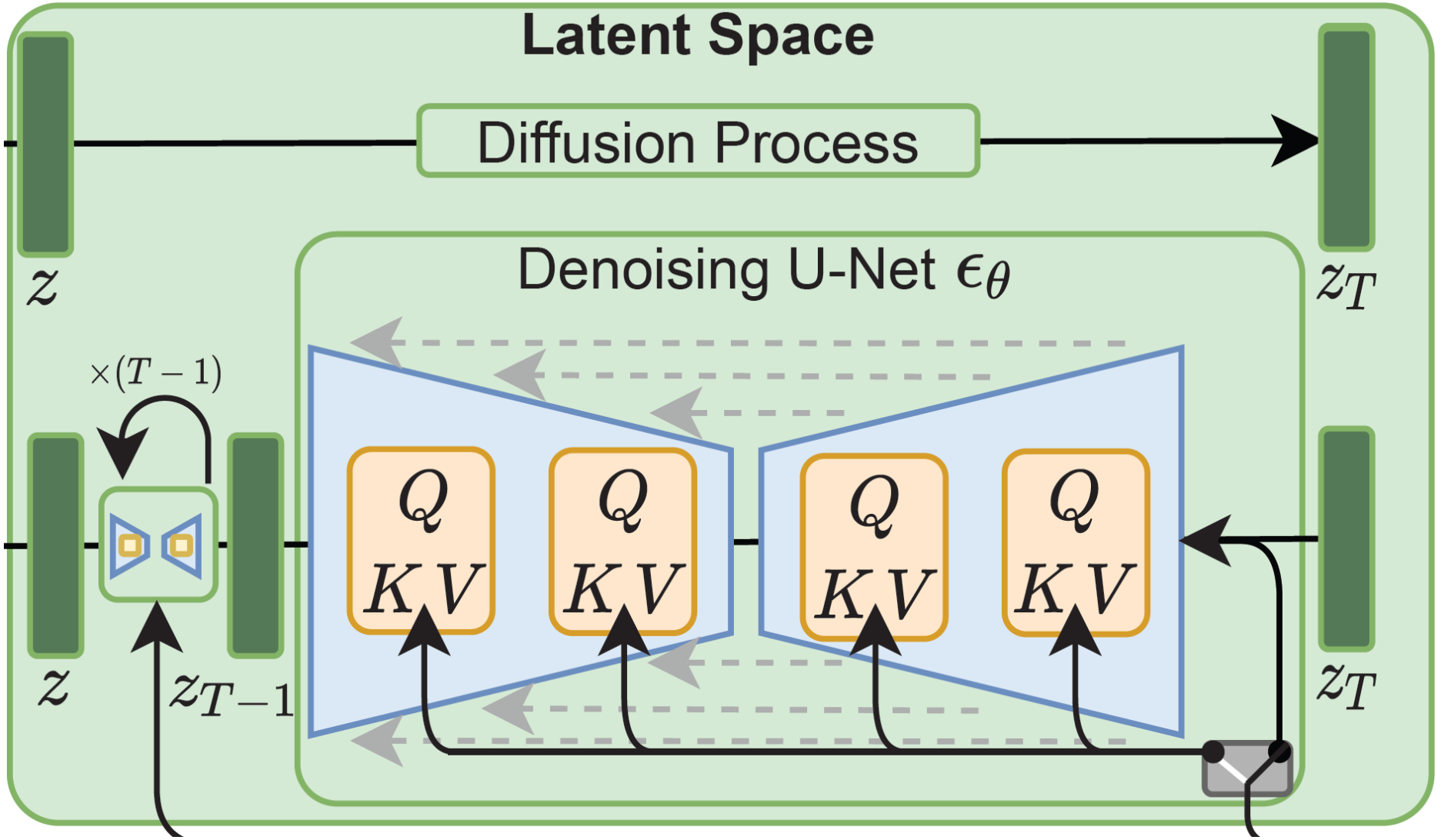


$$f = H/h = W/w.$$

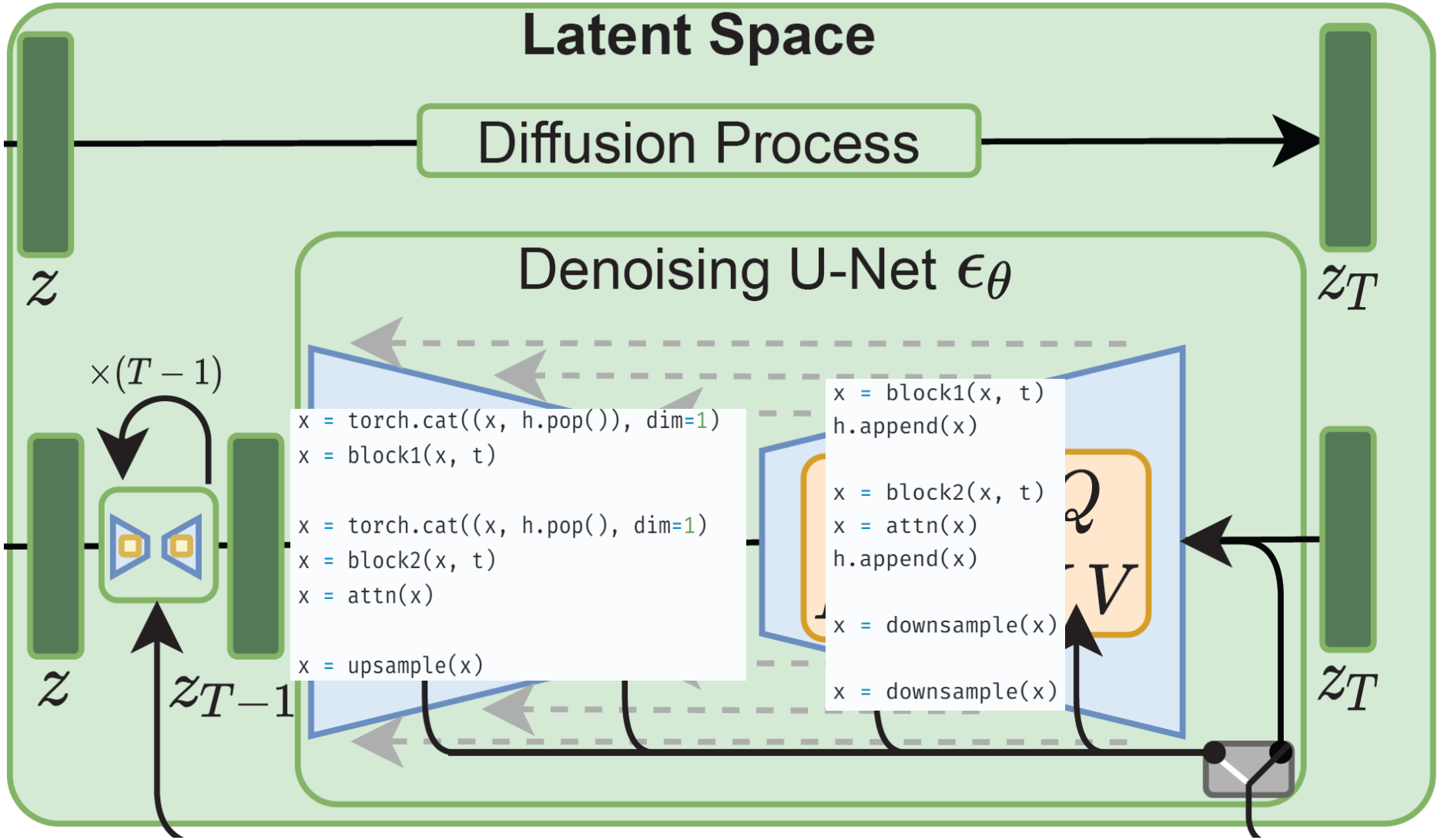
$$f = 2^m$$

The encoder and decoder can either be trained or use pretrained models.

03. Latent Diffusion Model

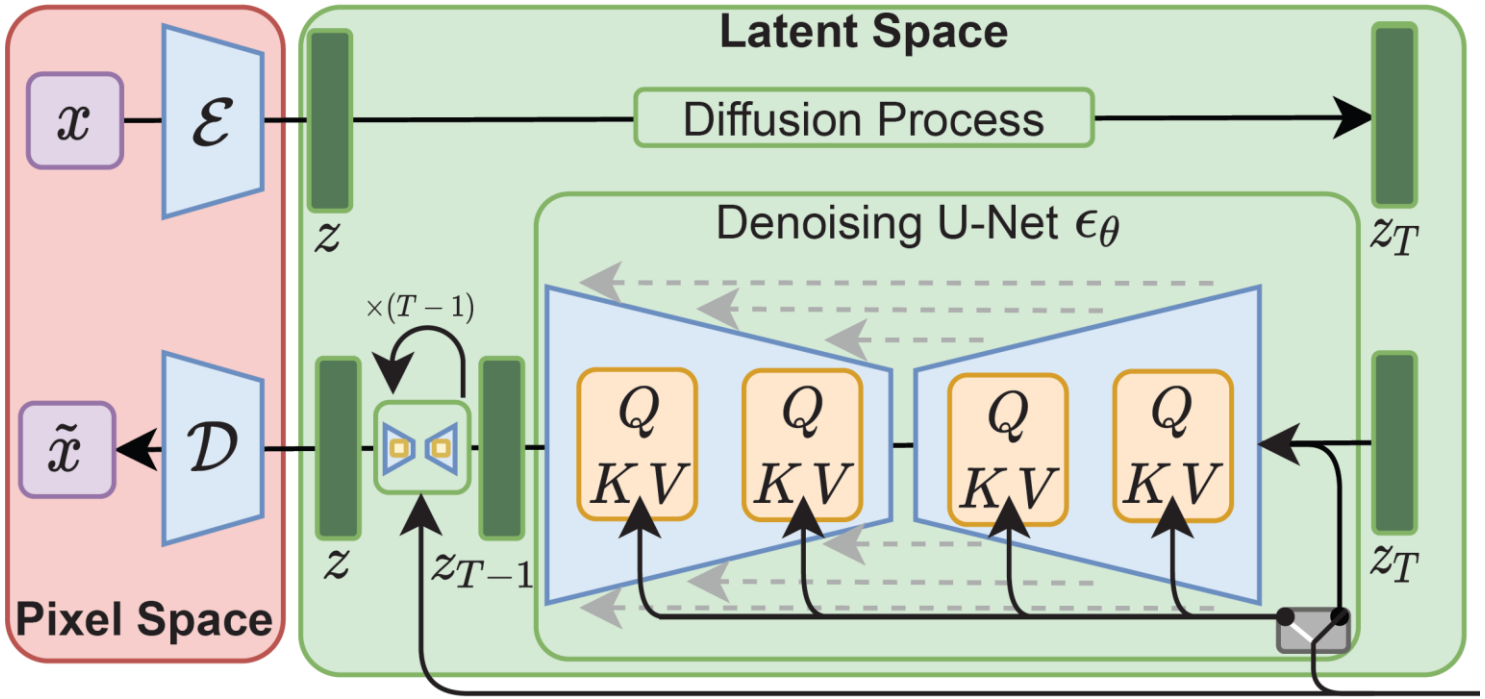


03. Latent Diffusion Model



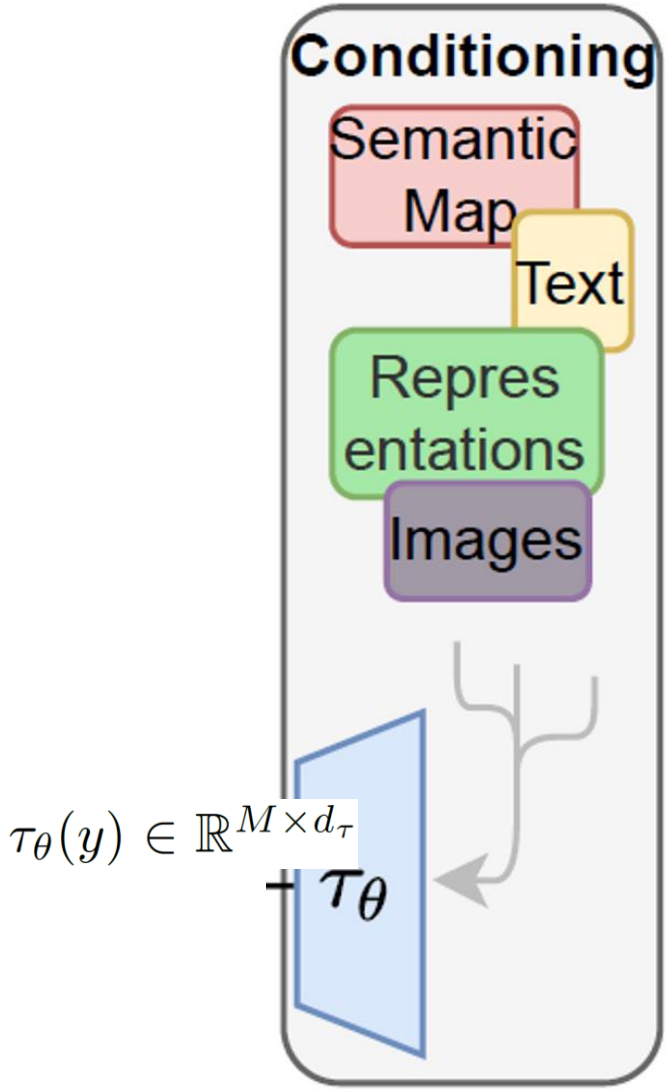
https://velog.io/@yeomjinseop/DDPM-%EA%B5%AC%ED%98%84%ED%95%98%EA%B8%B03

03. Latent Diffusion Model



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

04. Conditioning



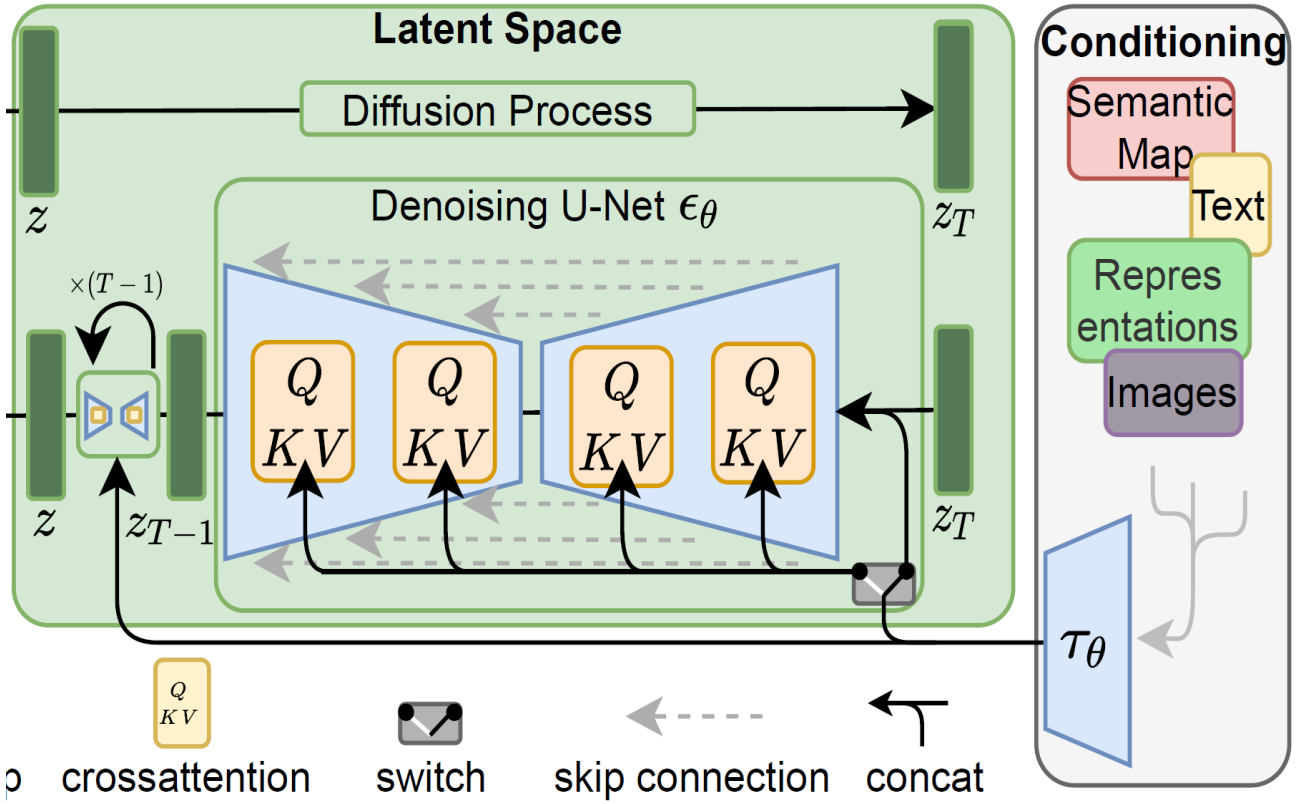
04. Conditioning

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

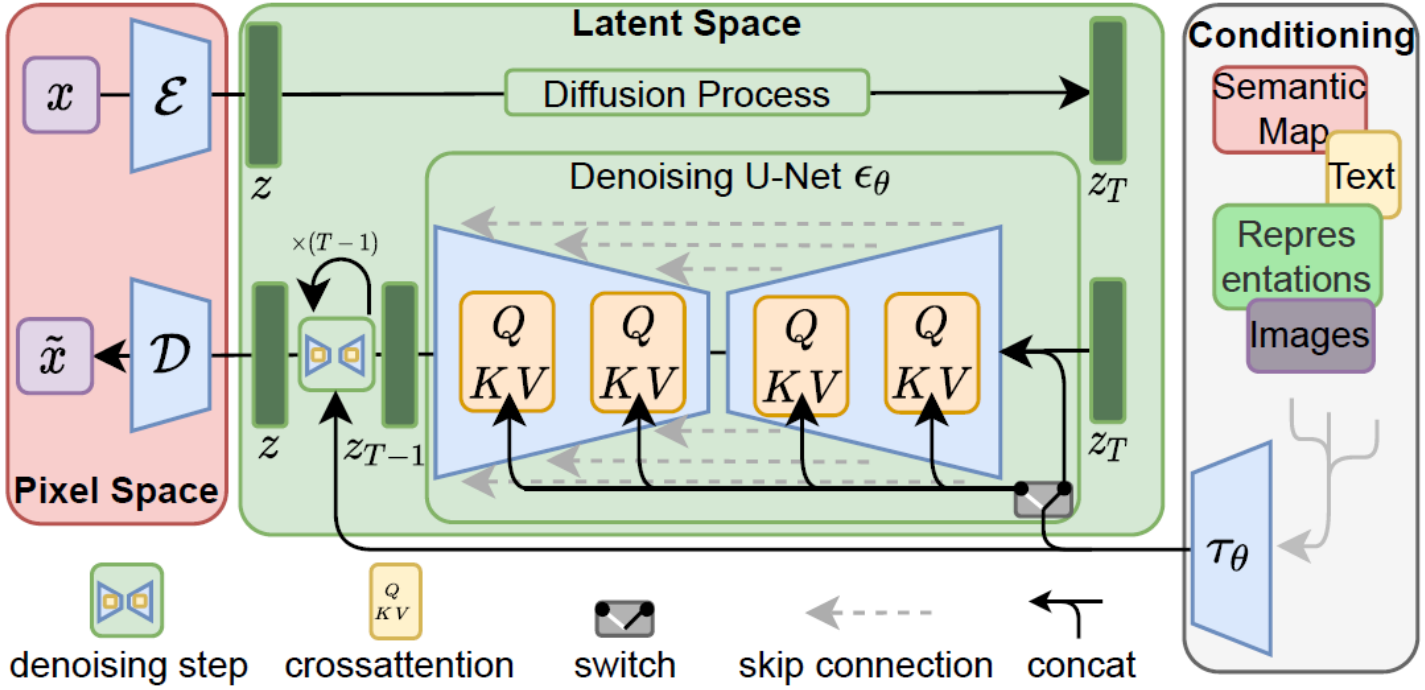
$$Q = W_Q^{(i)} \cdot \phi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\epsilon(y)$$

$$\phi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$$

$$W_V^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}, W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}, W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$$

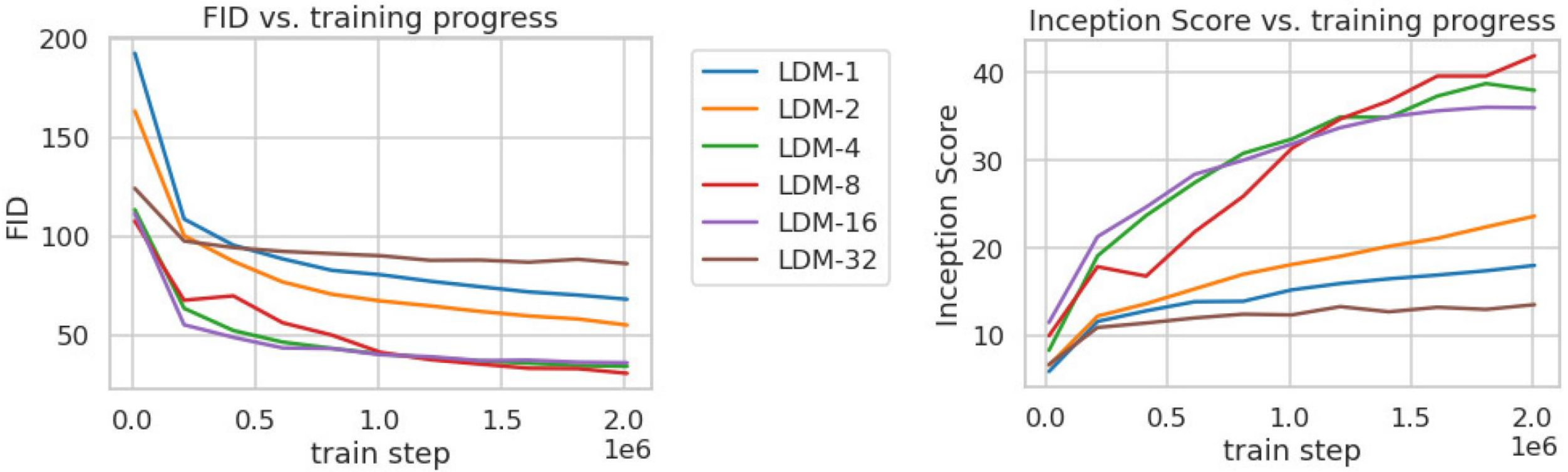


04. Conditioning



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

05. Experiment



05. Experiment

Text-Conditional Image Synthesis				
Method	FID ↓	IS↑	Nparams	
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] $s = 1.5$

06. Limitation

- While LDM significantly reduces computational requirements compared to pixel-based methods, sampling remains slower than GANs.
- Additionally, the use of LDMs is questionable when high precision is required.

More.....

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis

Dustin Podell **Zion English** **Kyle Lacey** **Andreas Blattmann** **Tim Dockhorn**

Jonas Müller **Joe Penna** **Robin Rombach**

Stability AI, Applied Research

Code: `https://github.com/Stability-AI/generative-models` *Model weights*: `https://huggingface.co/stabilityai/`

End