Low-Rank Adaptation / Stable Diffusion / LLaMa

24. 10. 31.

Mingyu Kim

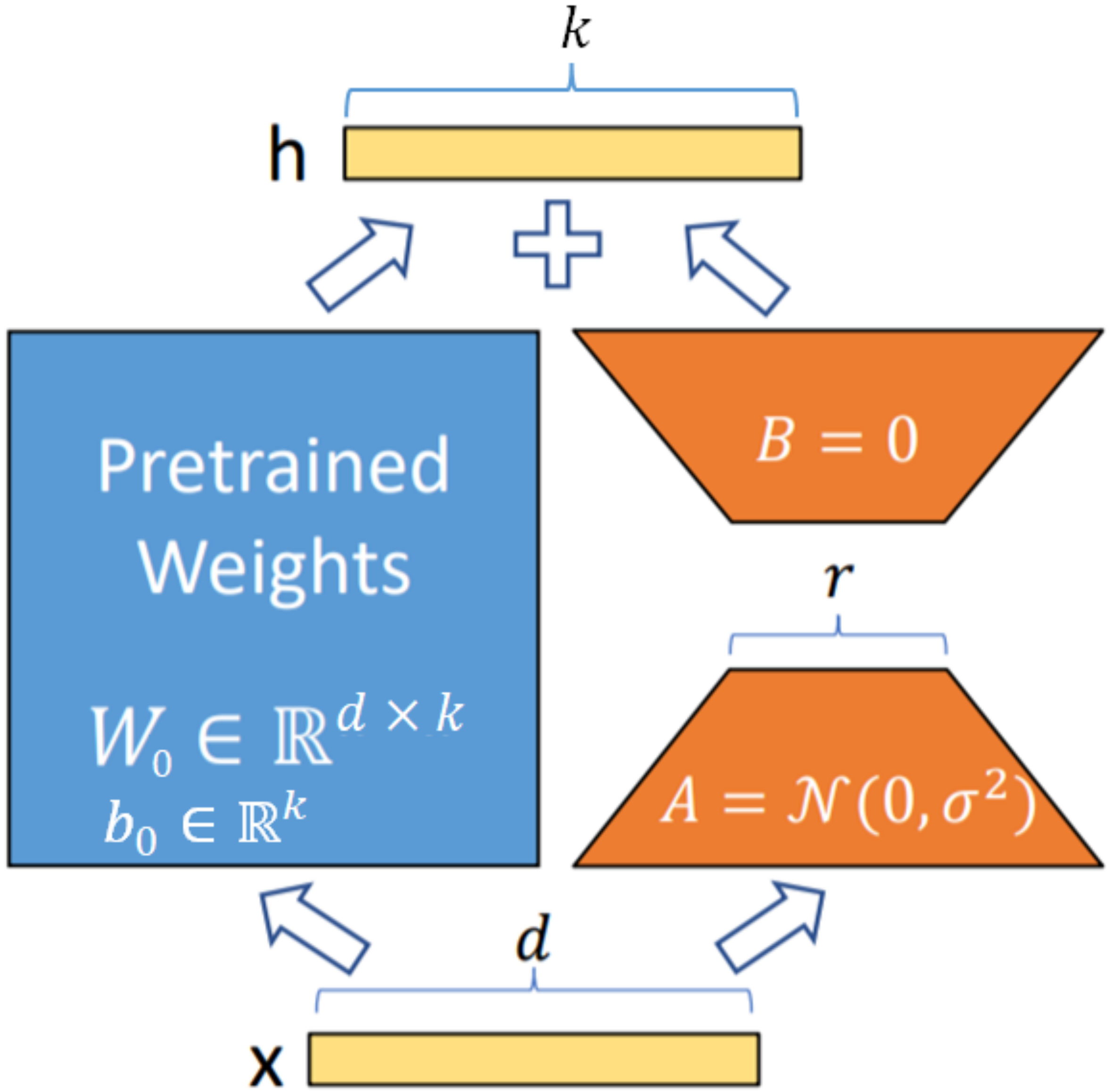
24.10.31

Repository

https://github.com/MingyuKim87/day5.git

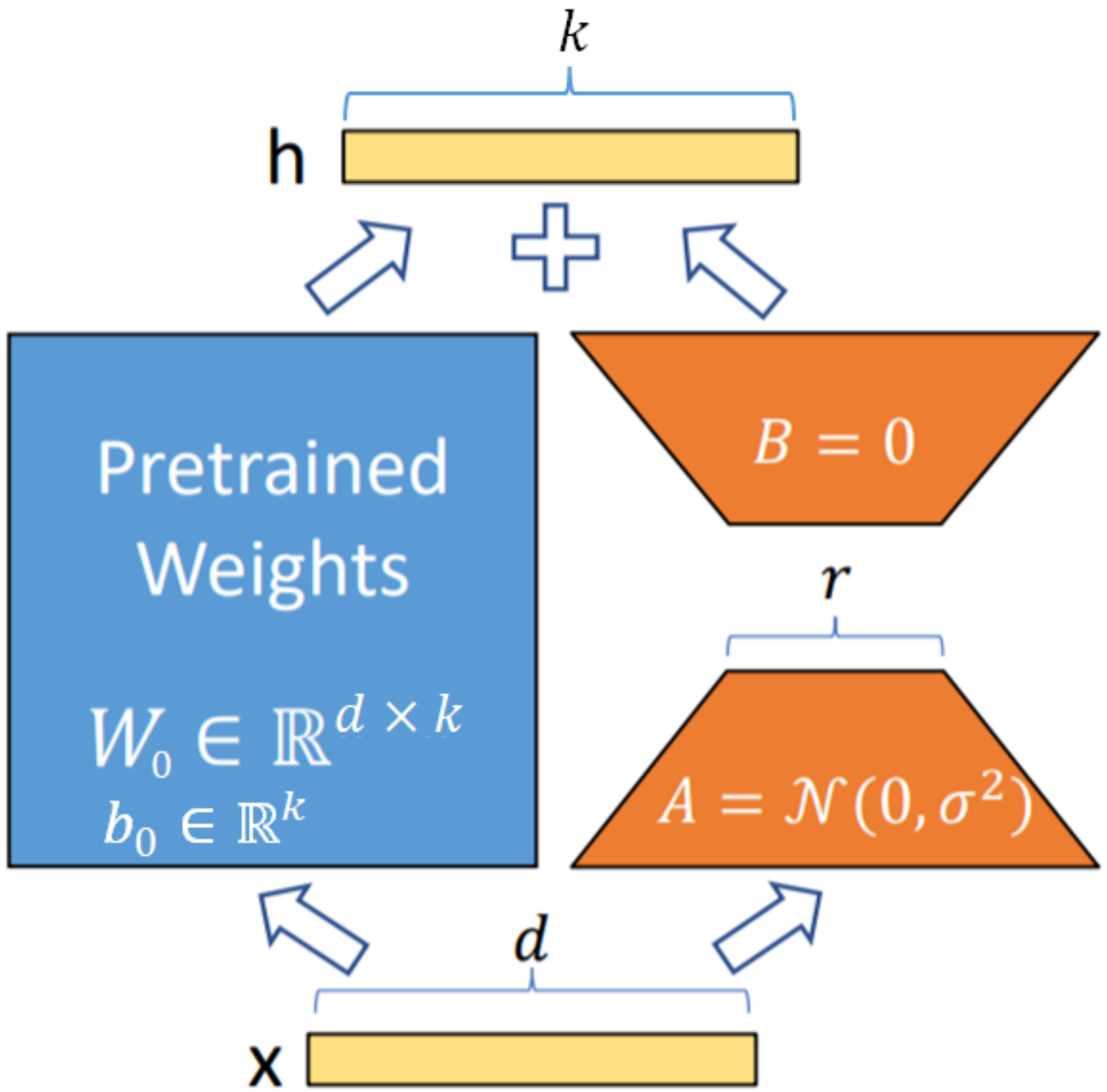
Low-Rank Adaptation

- The Pre-trained Network: $h = W_0 x + b_0$ where $W_0 \in \mathbb{R}^{d \times d}$, $b_0 \in \mathbb{R}^{d \times 1}$
- The LoRA Network : A, B where $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times d}$ and $d \ll \min(d, r)$
  - $A \sim N(0, \sigma^2)$, $B = 0$
- $h = (W_0 x + b_0) + sBAx = W_0 x + s\Delta W x = (W_0 + s\Delta W)x + b_0$ where $s = \alpha / r$ (α : hyper-parameter)



Low-Rank Adaptation

- For instance, $d = 768, k = 256 \rightarrow$ the number of parameters W_0 : 196,608
  - $A : 767 \times 4 = 3{,}068$
  - $B : 4 \times 256 = 1{,}024$
- $A + B = 4{,}092$



24.10.31

Stable Diffusion V1

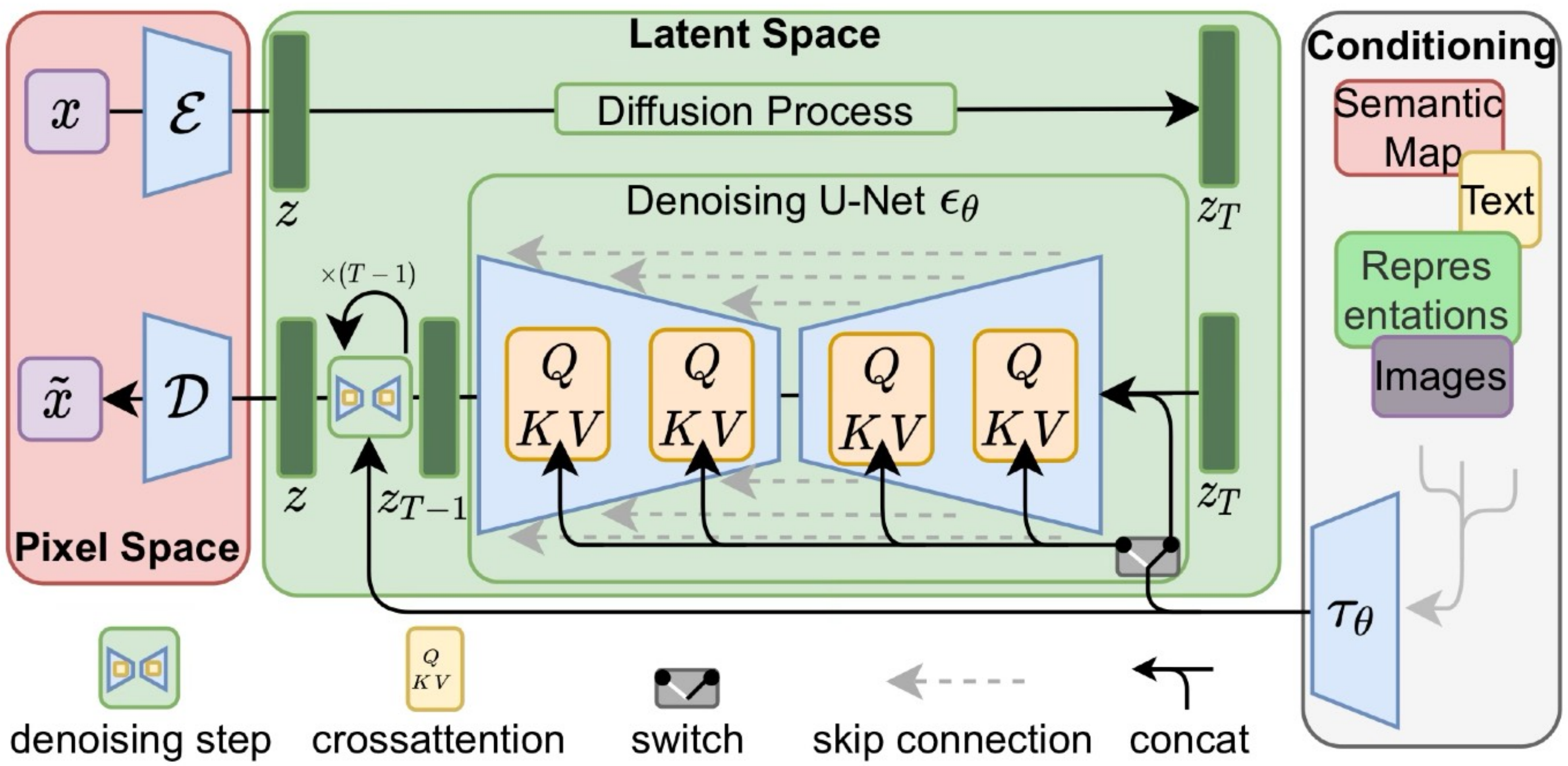
- Limitations of Previous Diffusion Models
  - Earlier diffusion models used a pixel-space UNet framework.
  - Generating high-resolution images with these models faces significant computational challenges due to the large amount of pixel data.

Model (*reg.*-type)	train throughput samples/sec.	sampling throughput[†] @256	@512	train+val hours/epoch	FID@2k epoch 6
LDM-1 (no first stage)	0.11	0.26	0.07	20.66	24.74
LDM-4 (*KL*, w/ attn)	0.32	0.97	0.34	7.66	15.21
LDM-4 (*VQ*, w/ attn)	0.33	0.97	0.34	7.04	14.99
LDM-4 (*VQ*, w/o attn)	0.35	0.99	0.36	6.66	15.95

Table 6. Assessing inpainting efficiency. [†]: Deviations from Fig. 7 due to varying GPU settings/batch sizes *cf*. the supplement.

Stable Diffusion V1

- Advancements with Stable Diffusion
  - Stable Diffusion introduces diffusion models that operate within *latent spaces*, rather than direct pixel-space.
  - This approach is made possible by employing a Variational Autoencoder (VAE) architecture, which enables efficient high-resolution image synthesis by encoding images into compact latent representations.



24.10.31

Stable Diffusion V1

- Diversity in Conditional Information Sources
  - Text Information: Allows for semantic input, guiding image generation with descriptive text.
  - Sketch Information: Adds structure or outlines to provide foundational visual guidance.
  - Scene Generation: Combines various conditional inputs to produce cohesive and context-rich scenes.

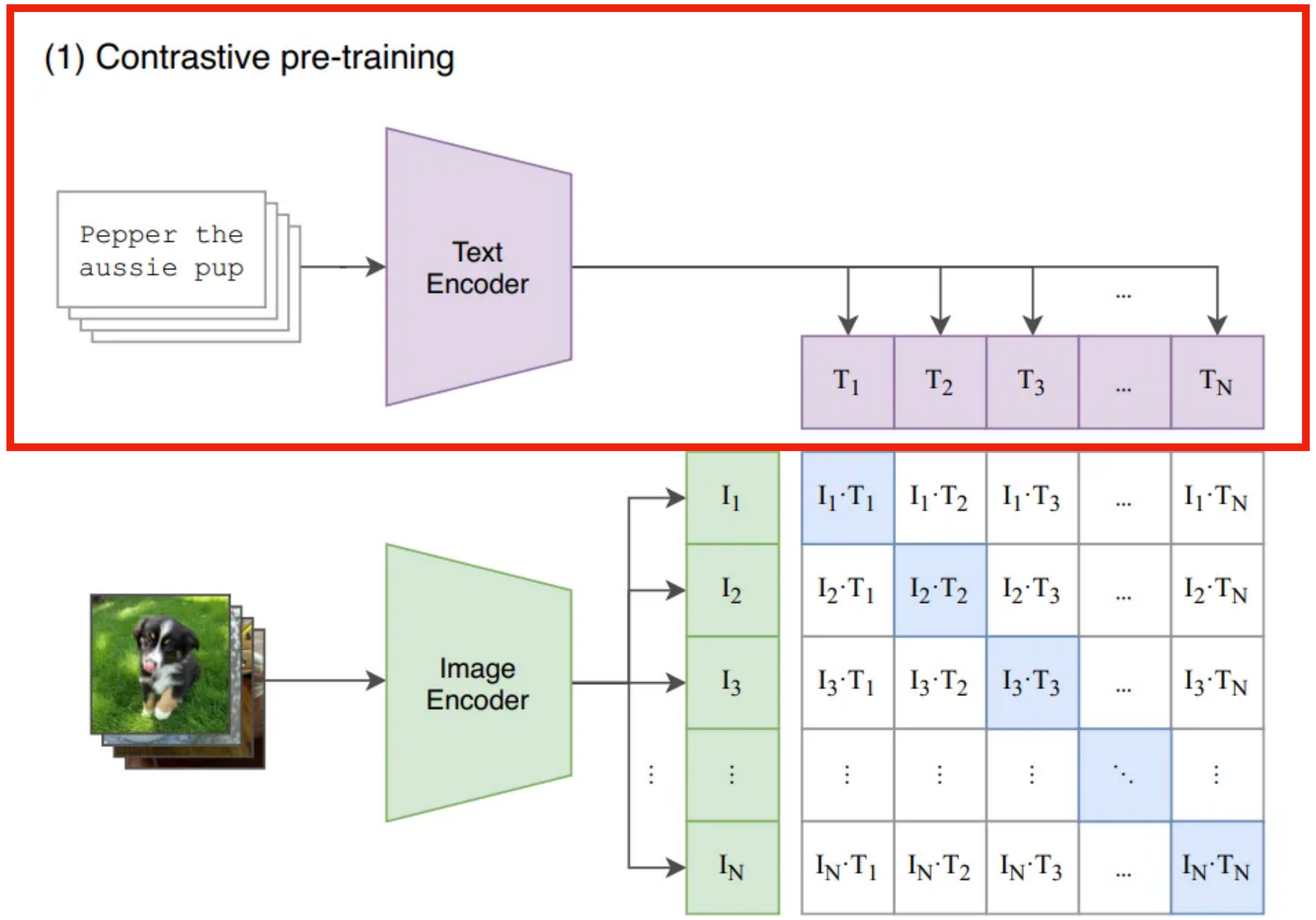
Figure 8. Layout-to-image synthesis with an *LDM* on COCO [4], see Sec. 4.3.1. Quantitative evaluation in the supplement D.3.

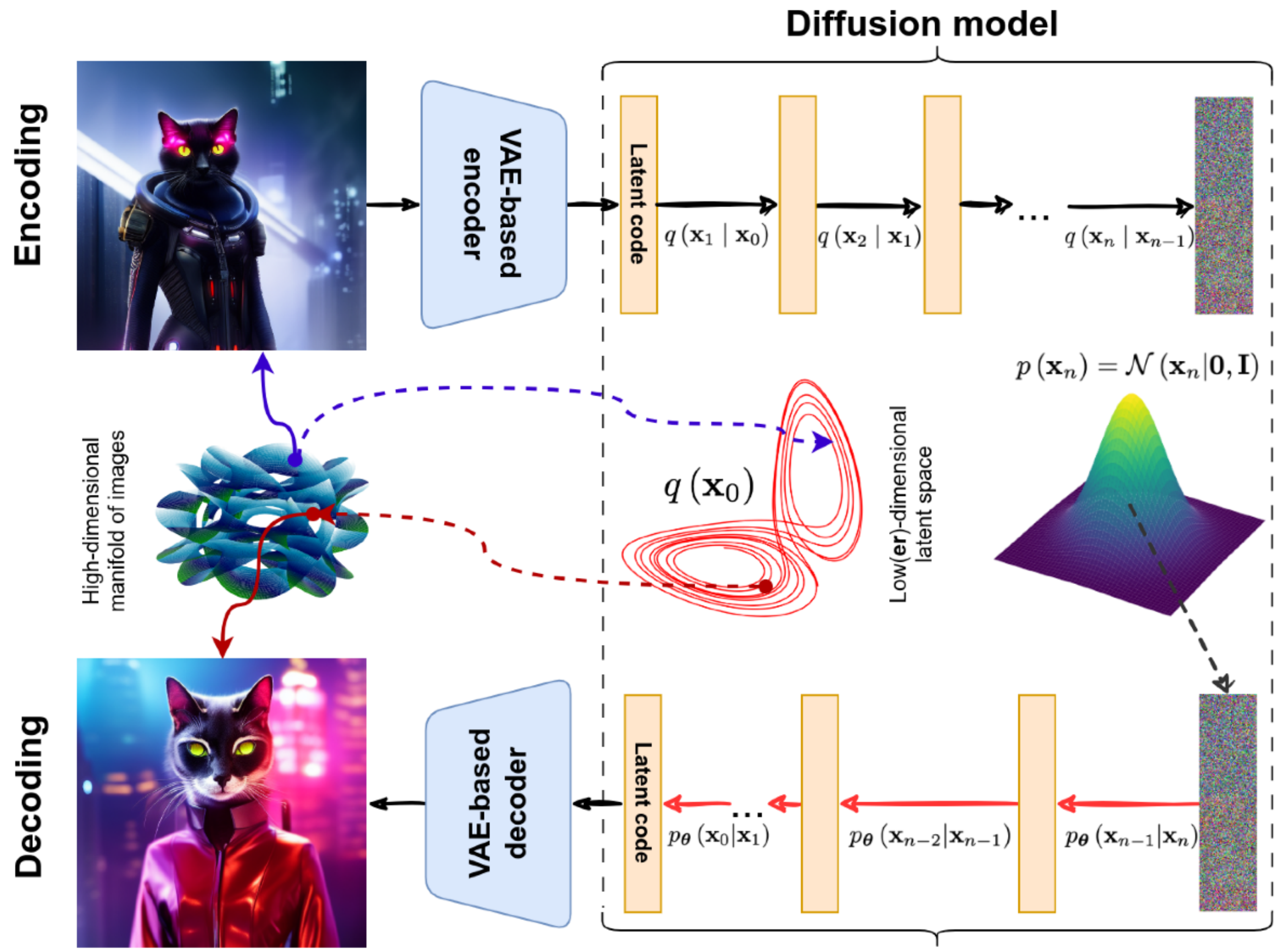
Figure 9. A *LDM* trained on 256^2 resolution can generalize to larger resolution (here: 512×1024) for spatially conditioned tasks such as semantic synthesis of landscape images. See Sec. 4.3.2.

24.10.31

Stable Diffusion V1

- Text-to-Image Architecture
  - CLIP Text Encoder (+ Tokenizer)
  - VAE: Encoder / Decoder
  - UNET: Diffusion Models





24.10.31

Guidance

- Three Equivalent Interpretations

24.10.31

Guidance

- Tweedie’s Forumula

Guidance

- Score-Based Generative Models

24.10.31

Guidance

- Classifier Guidance

$$\nabla \log p(\boldsymbol{x}_t|y) = \nabla \log\left(\frac{p(\boldsymbol{x}_t)p(y|\boldsymbol{x}_t)}{p(y)}\right) \tag{163}$$

$$= \nabla \log p(\boldsymbol{x}_t) + \nabla \log p(y|\boldsymbol{x}_t) - \nabla \log p(y) \tag{164}$$

$$= \underbrace{\nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}} + \underbrace{\nabla \log p(y|\boldsymbol{x}_t)}_{\text{adversarial gradient}} \tag{165}$$

  - Implementation

$$\nabla \log p(\boldsymbol{x}_t|y) = \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(y|\boldsymbol{x}_t) \tag{166}$$

Guidance

- Classifier Free Guidance
  - We assume that $\log p(x_t \mid y)$ is accessible

$$\begin{aligned}
\nabla \log p(\boldsymbol{x}_t|y) &= \nabla \log \left(\frac{p(\boldsymbol{x}_t)p(y|\boldsymbol{x}_t)}{p(y)} \right) && (163)\\
&= \nabla \log p(\boldsymbol{x}_t) + \nabla \log p(y|\boldsymbol{x}_t) - \nabla \log p(y) && (164)\\
&= \underbrace{\nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}} + \underbrace{\nabla \log p(y|\boldsymbol{x}_t)}_{\text{adversarial gradient}} && (165)
\end{aligned}$$

  - Adversarial gradient breaks down:

$$\nabla \log p(y|\boldsymbol{x}_t) = \nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t) \qquad (167)$$

  - Substituting this into:

$$\begin{aligned}
\nabla \log p(\boldsymbol{x}_t|y) &= \nabla \log p(\boldsymbol{x}_t) + \gamma \left(\nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t) \right) && (168)\\
&= \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(\boldsymbol{x}_t|y) - \gamma \nabla \log p(\boldsymbol{x}_t) && (169)\\
&= \underbrace{\gamma \nabla \log p(\boldsymbol{x}_t|y)}_{\text{conditional score}} + \underbrace{(1-\gamma) \nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}} && (170)
\end{aligned}$$

24.10.31

Stable Diffusion V1

- Classifier Free Guidance outperforms other methods and its simple generations.
  - FID (Fréchet Inception Distance): how close the generated images are to real images by comparing the distributions of features
  - IS (Inception Score): Represent Objects + Diverse Across Classes

Text-Conditional Image Synthesis				
Method	FID $\downarrow$	IS$\uparrow$	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] $s = 1.5$

Q&A

End