西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Adding Conditional Control to Text-to-Image Diffusion Models

(ControlNet)

ICCV2023 best paper

Lvmin Zhang and Maneesh Agrawala, Stanford University

主讲人：阮皓

2024.04.10

About the Author

Figure 1. Lvmin's reading a Twitter meme after waking up at 7:30 am. © *Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).*

As shown in Fig. 1, Lvmin Zhang (Lyumin Zhang) is a Ph.D. student in Computer Science advised by Prof. Maneesh Agrawala at Stanford University since 2022. Before that, he was a Research Assistant in the lab of Prof. Tien-Tsin Wong at the Chinese University of Hong Kong since 2021. He has also collaborated with Prof. Edgar Simo-Serra on many interesting projects. He received his bachelor's degree of B.Eng. from Soochow University in 2021, supervised by Prof. Yi Ji and Prof. Chunping Liu.

Lvmin's research fields include computational art and design, interactive content creation, computer graphics, image and video processing, and **ANIME**! He loves these and organized a special interest research group called Style2Paints Research. He also developed an anime drawing software called Style2Paints.

You might be interested in:

- Lvmin's researches and publications in his research page.
- Lvmin's projects in his GitHub profile.
- Lvmin's email addresses: lvmin AT cs.stanford.edu / lvminzhang AT acm.org / lvminzhang AT siggraph.org.

In the leisure time Lvmin likes developing games. Lvmin is the author of an Unity card game called YGOPro2. If you search this game in Google or YouTube, you will find it popular. This game has been translated to many languages and has fans all over the world.

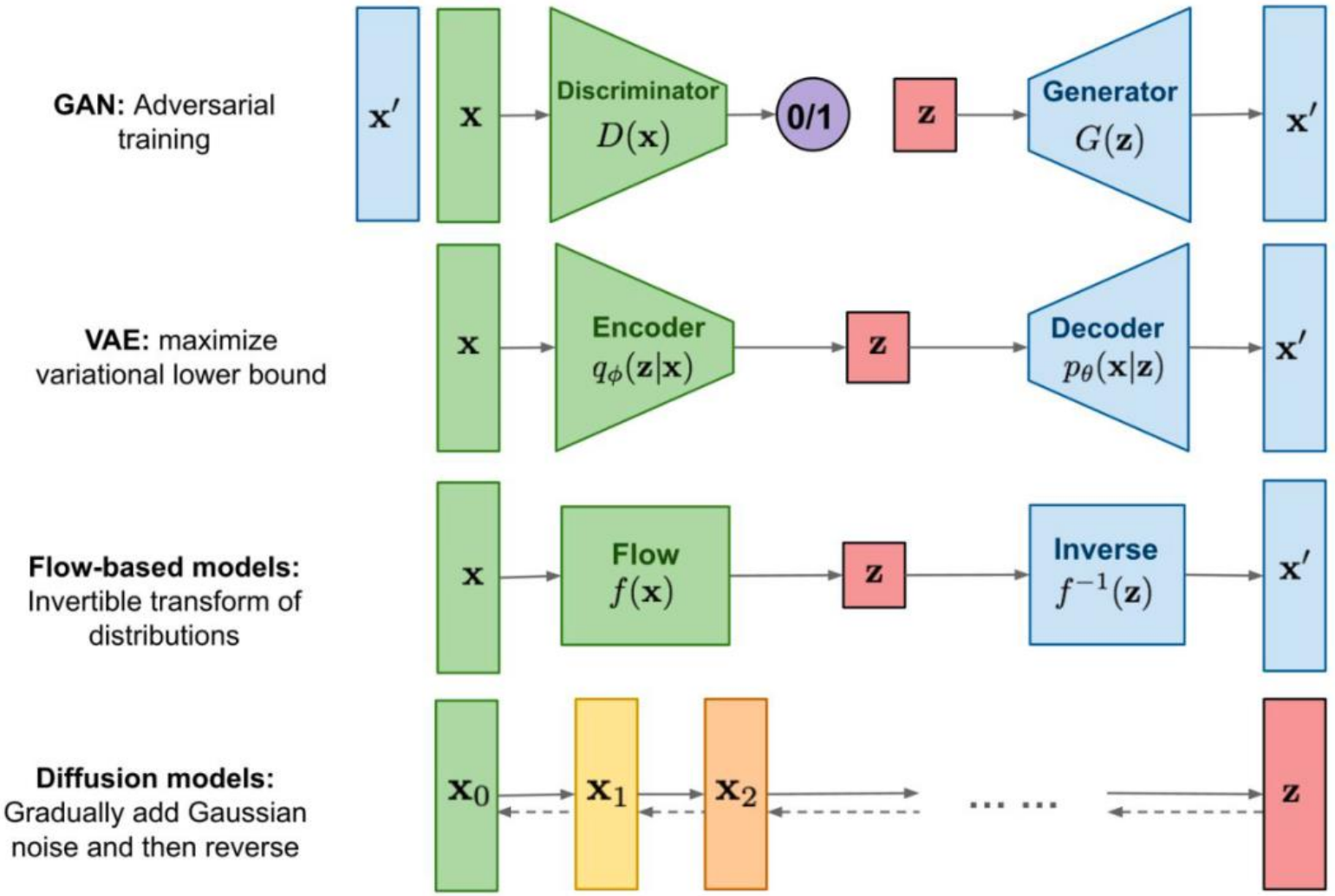
About

Let us control diffusion models!

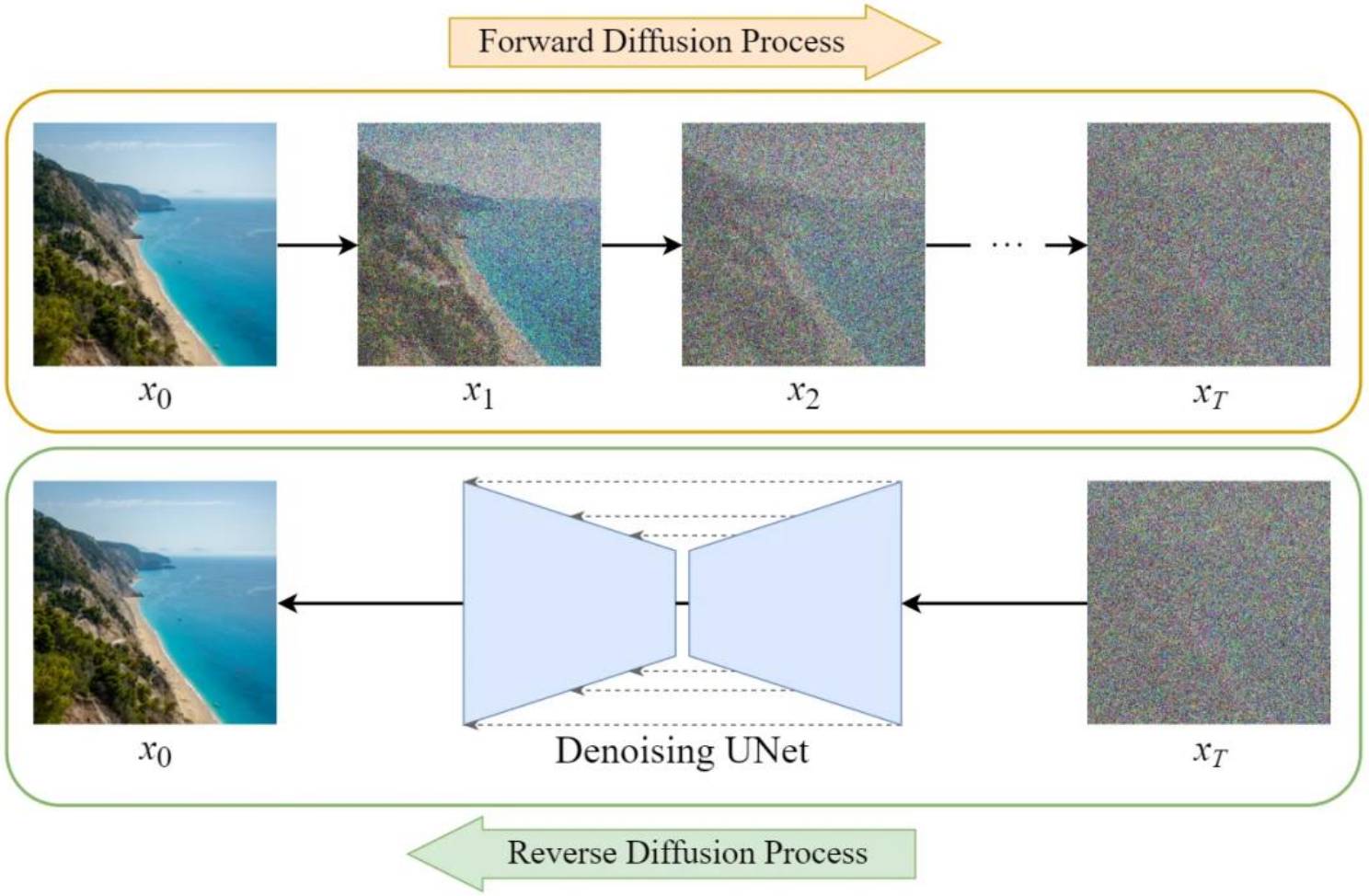
- Readme
- Apache-2.0 license
- Activity
- 27.5k stars
- 208 watching
- 2.5k forks

Report repository

Review——Generative(Text2img) Model



Review——Diffusion Model



1. 前向扩散过程（Forward Diffusion Process）

   向图片中添加噪声；

2. 反向扩散过程（Reverse Diffusion Process）

   去除图片中的噪声。

Review——Diffusion Model

For each training step:

1. Randomly select a time step & encode it

2. Add noise to image

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon$$

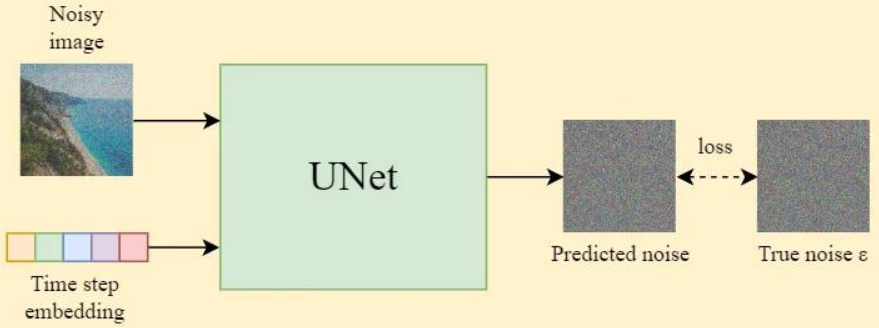
Adjust the amount of noise according to the time step t

$$\varepsilon \sim \mathcal{N}(0,1)$$

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$$

3. Train the UNet



[Denoising Diffusion Probabilistic Models (arxiv.org)]

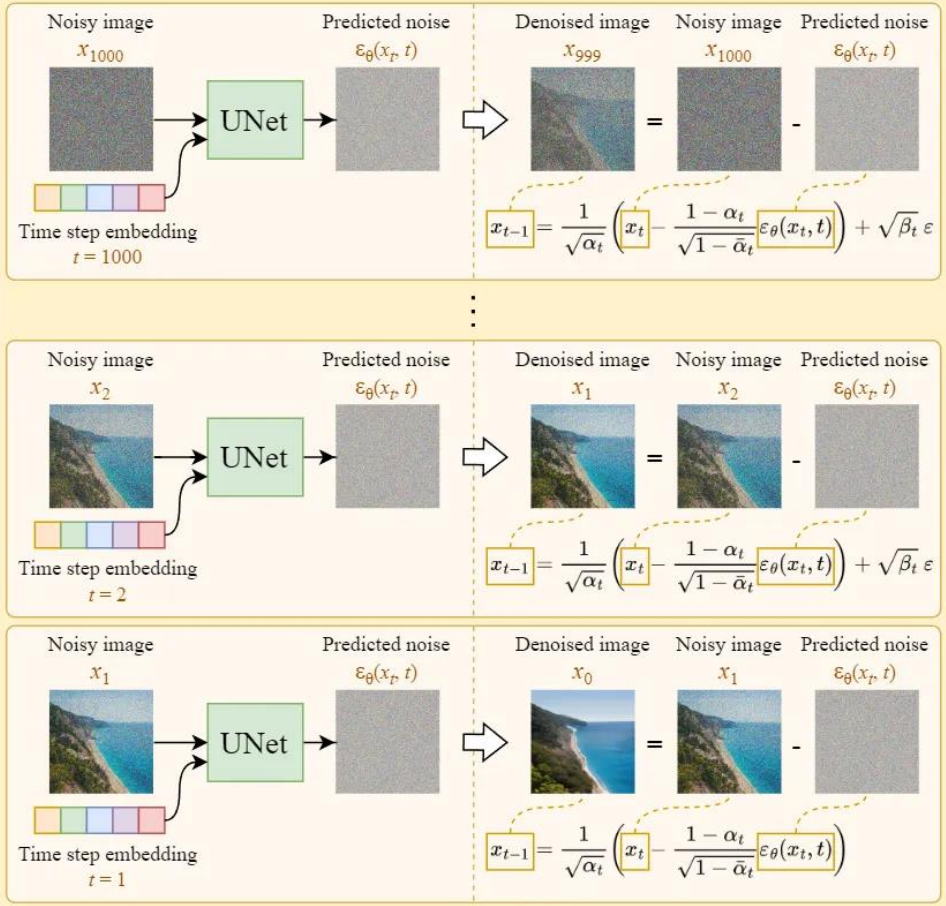
NeurIPS 2020

Reverse Diffusion / Denoising / Sampling

1. Sample a Gaussian noise

$x_T \sim N(0, \mathrm{I})$ E.g. $T = 1000$, $x_{1000} \sim N(0, \mathrm{I})$

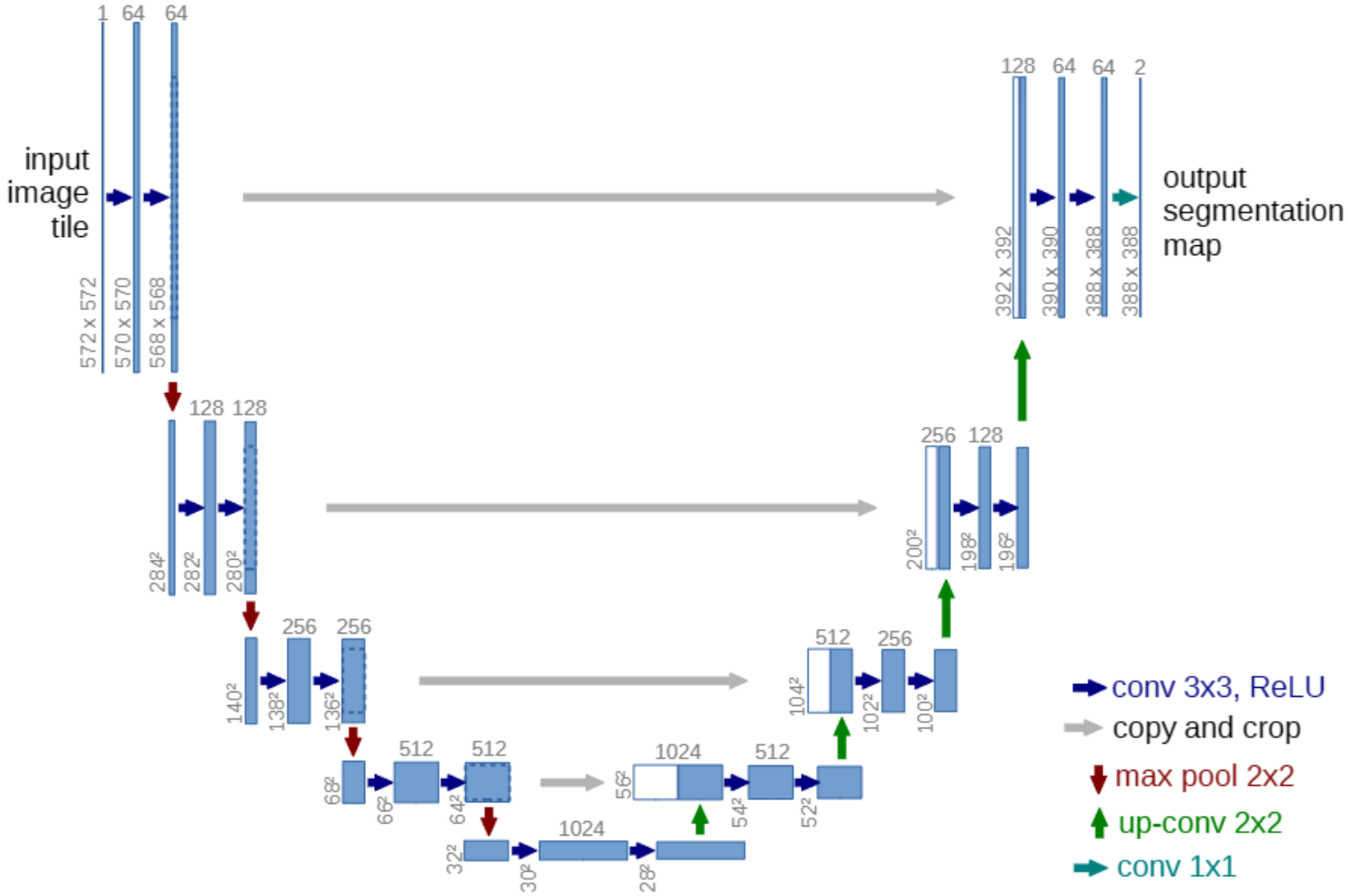
2. Iteratively denoise the image



3. Output the denoised image

Denoised image x_0

Review——Unet



U-Net: Convolutional Networks for Biomedical Image Segmentation (arxiv.org)

Miccai 2015

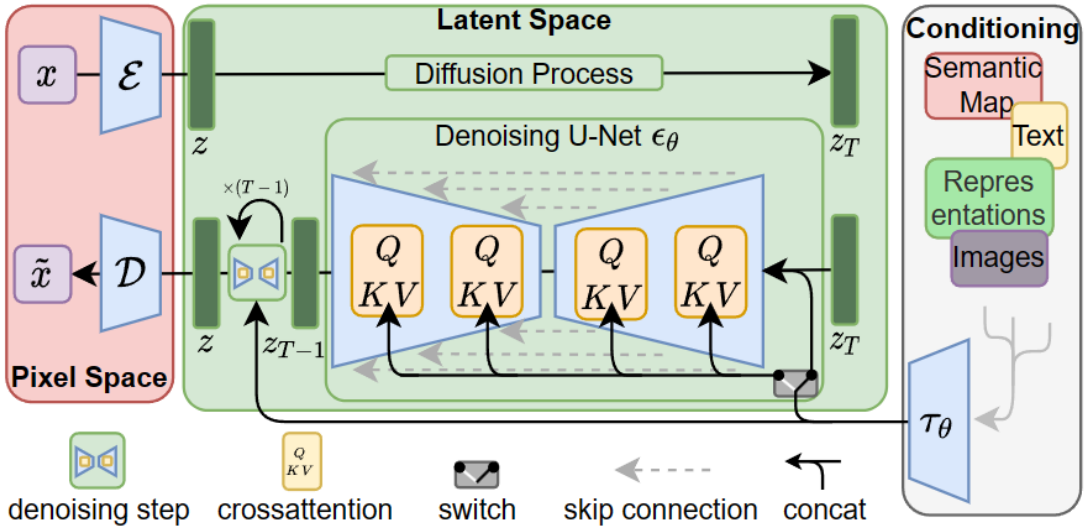
Review——Stable Diffusion

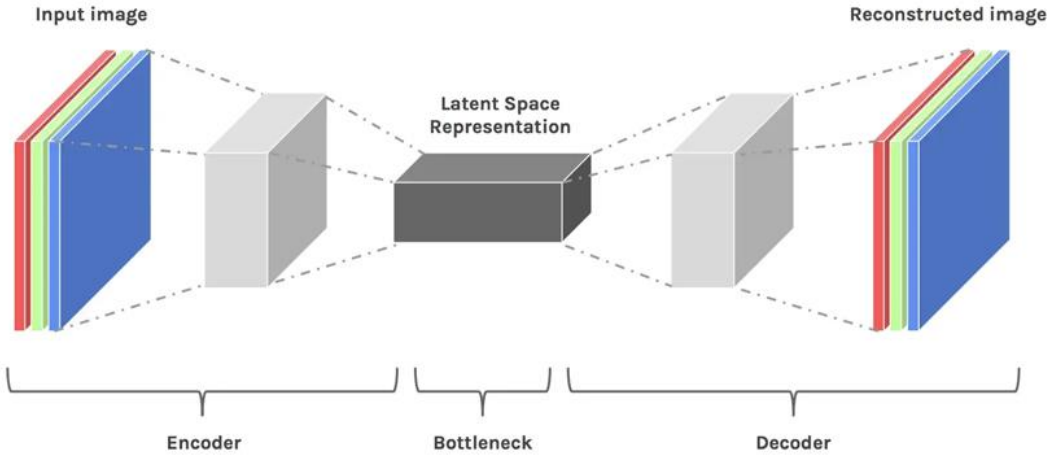
Diffusion的问题:

- 在反向扩散过程中需要把完整尺寸的图片输入到 U-Net，这使得当图片尺寸以及 time step t 足够大时，Diffusion 会非常的慢。
- 一张图片中的许多像素点可能与图片本身想要表达的内容是不相关的。
- 没有条件输入

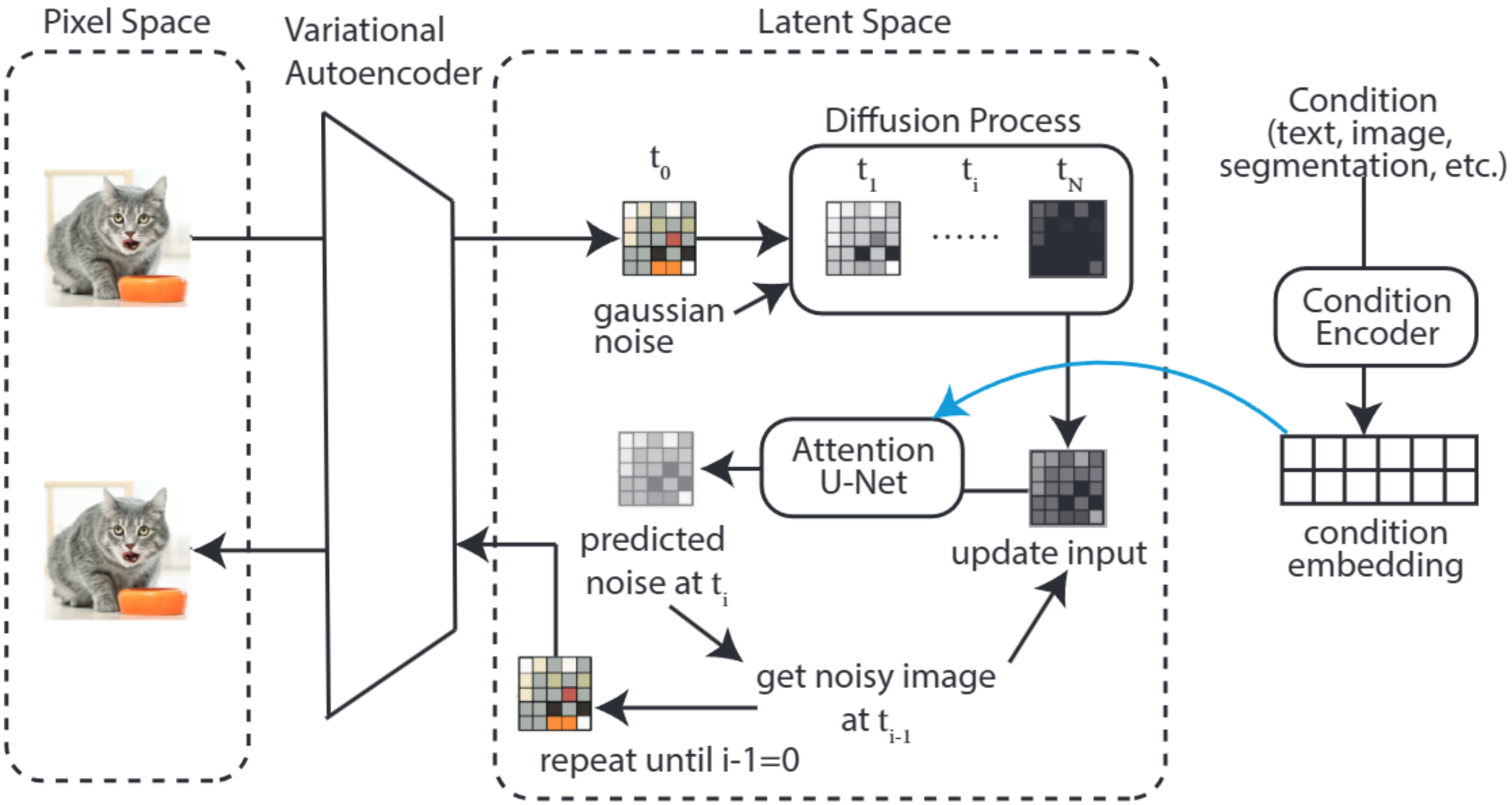
[High-Resolution Image Synthesis with Latent Diffusion Models (arxiv.org)]

CVPR 2022

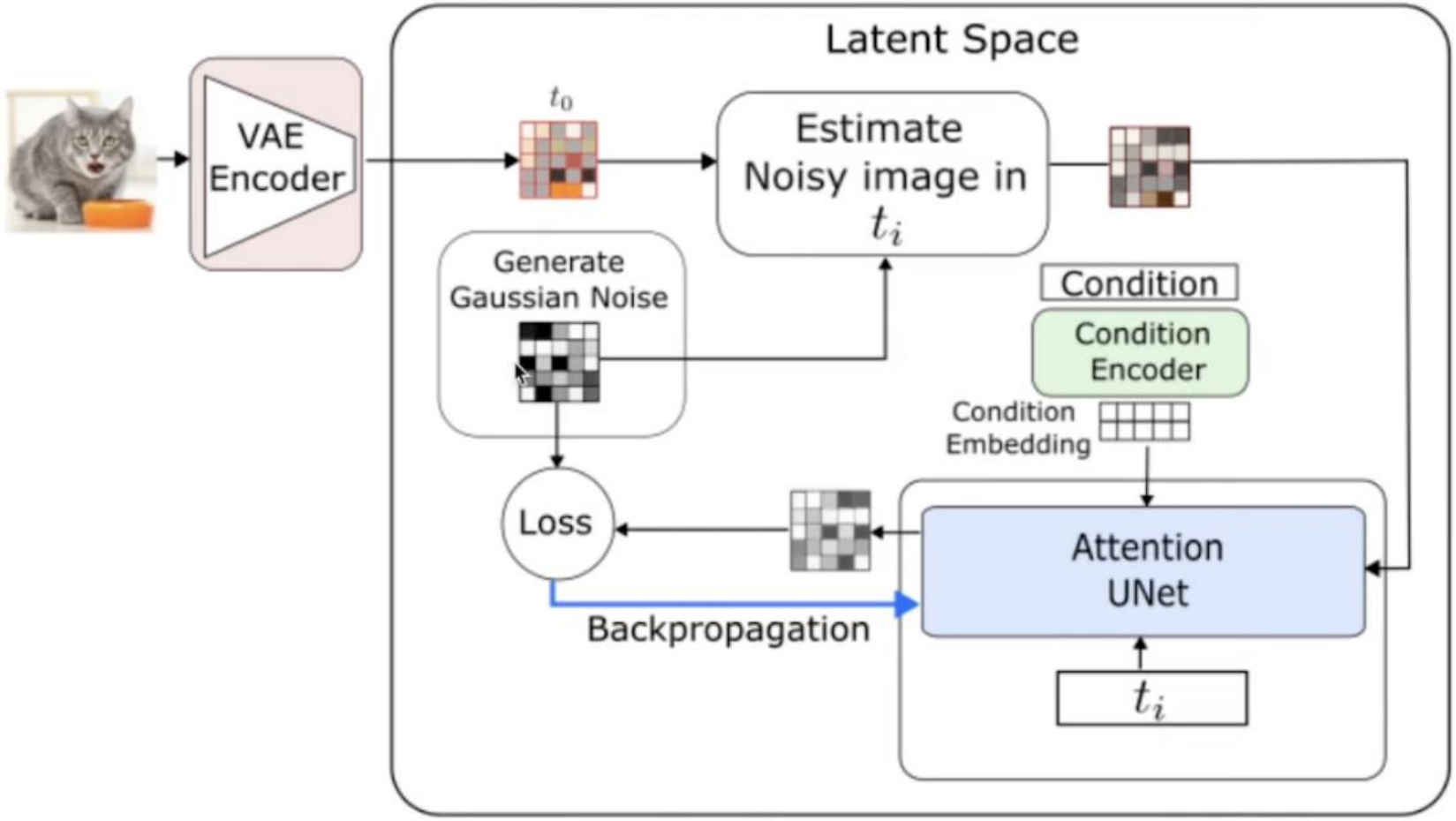




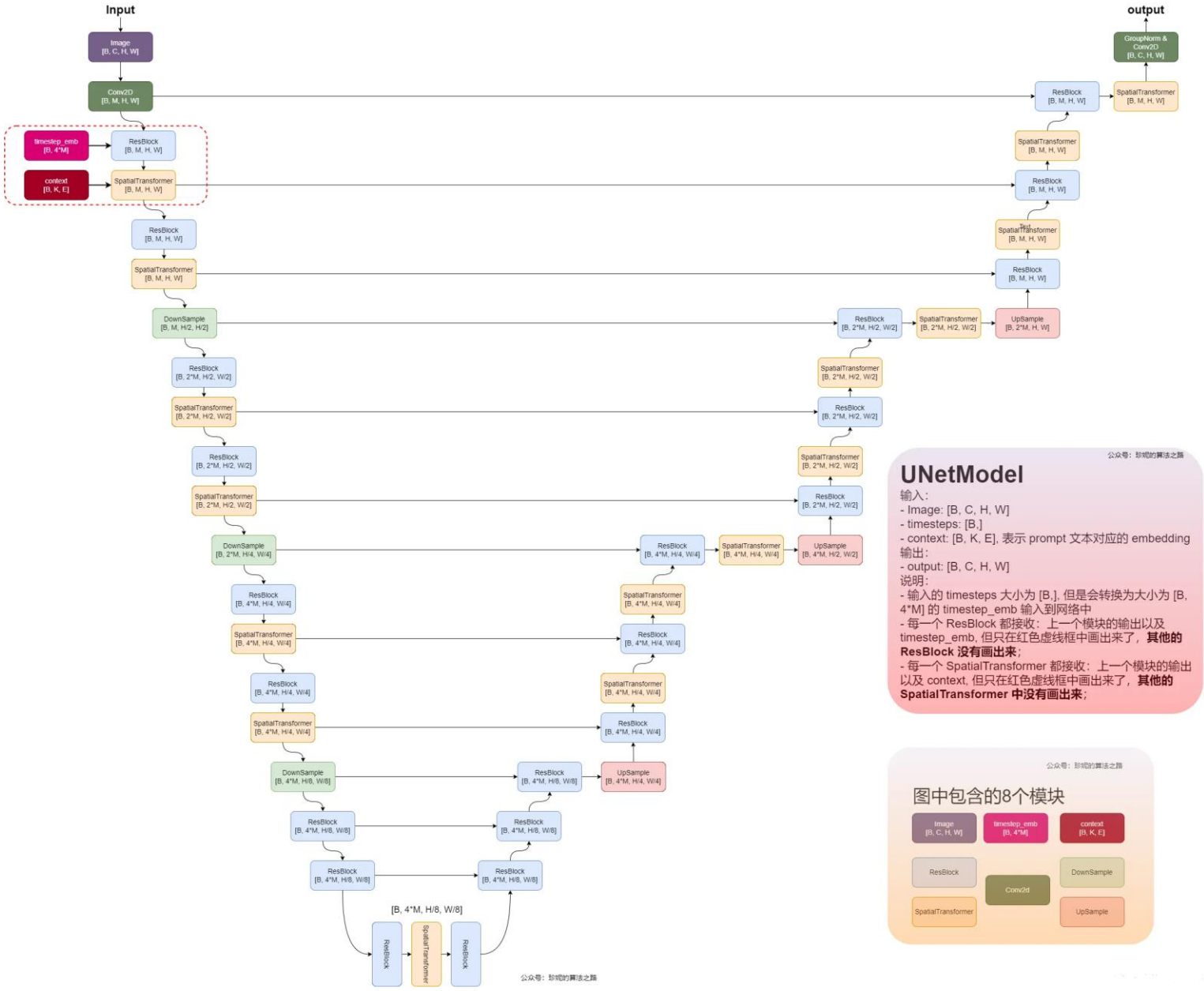
Review——Stable Diffusion



Review——Stable Diffusion



Review——Stable Diffusion



UNetModel

输入：
- Image: [B, C, H, W]
- timesteps: [B,]
- context: [B, K, E], 表示 prompt 文本对应的 embedding

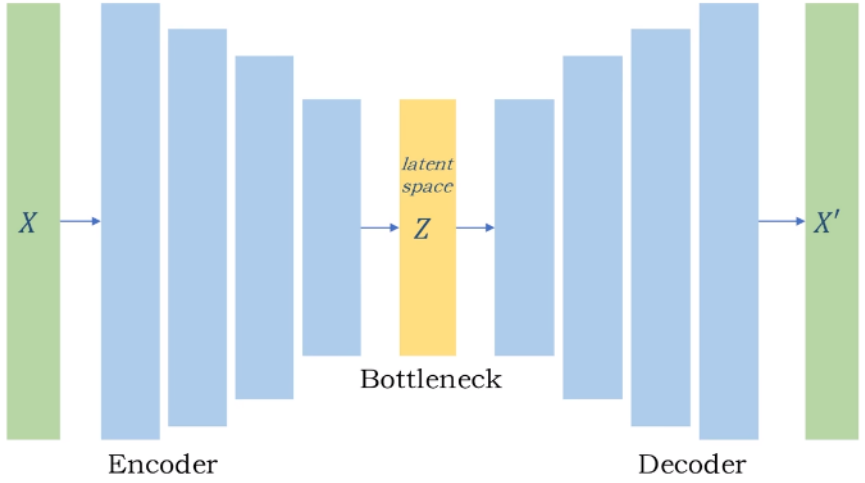
输出：
- output: [B, C, H, W]

说明：
- 输入的 timesteps 大小为 [B,], 但是会转换为大小为 [B, 4*M] 的 timestep_emb 输入到网络中
- 每一个 ResBlock 都接收：上一个模块的输出以及 timestep_emb, 但只在红色虚线框中画出来了，**其他的 ResBlock 没有画出来**；
- 每一个 SpatialTransformer 都接收：上一个模块的输出以及 context, 但只在红色虚线框中画出来了，**其他的 SpatialTransformer 中没有画出来**；

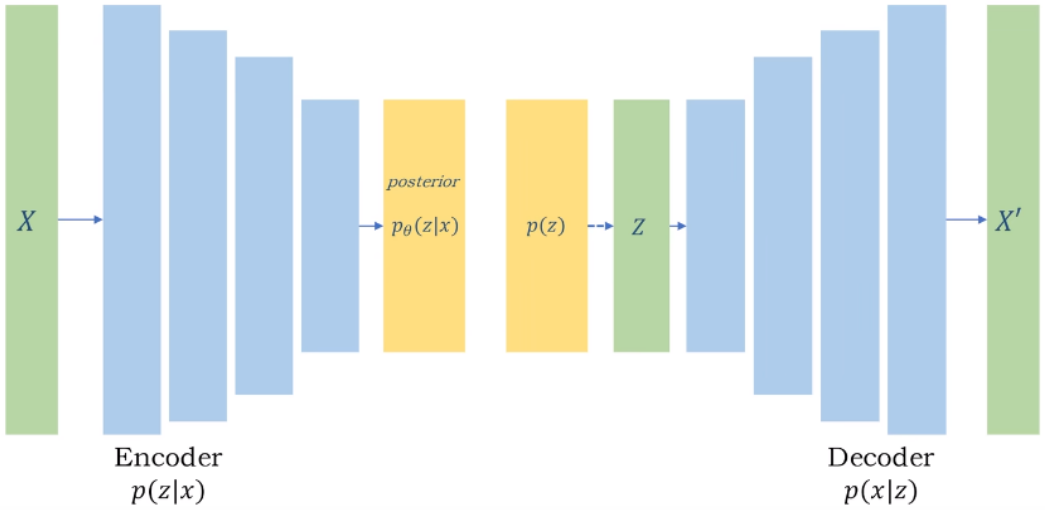
Review——AE，VAE

训练一个自动编码器模型来学习一个与原始图像空间感知上等价的潜在空间，可以在显著降低计算复杂度的同时，尽可能地保留图像的重要视觉信息和细节。

AutoEncoder



Variational AutoEncoder



AutoEncoder可以做生成任务吗?

$\{X_1, X_2, \ldots\ldots, X_n\}$

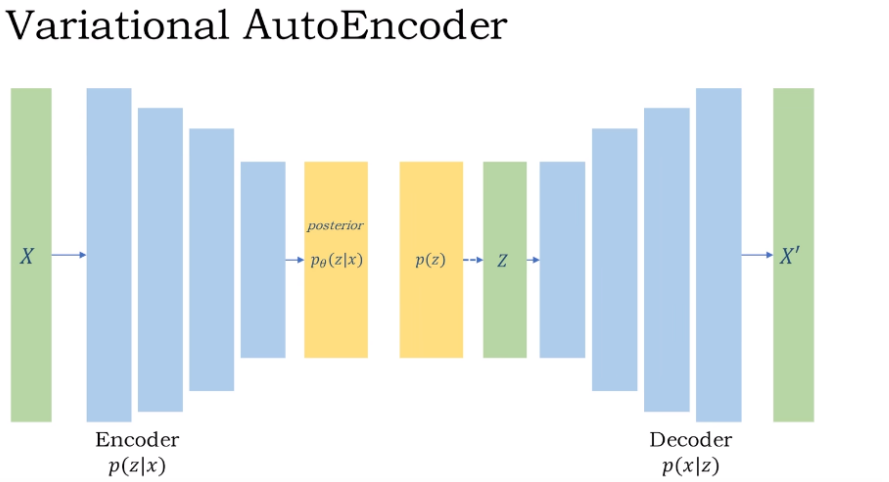
$\{Z_1, Z_2, \ldots\ldots, Z_n\}$

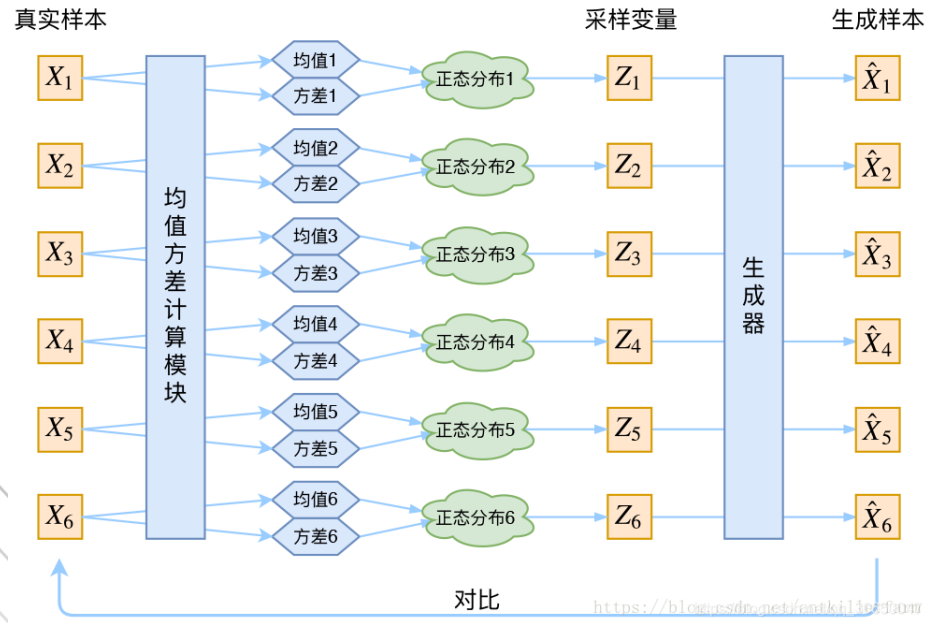
Z_{n+1}?

Auto-Encoding Variational Bayes (arxiv.org)

ICLR 2014

Review——AE，VAE





$$p(Z)\sim N(0,1)$$

$$p(\mathrm{Z}_i|X_i)\ \text{Known?}$$

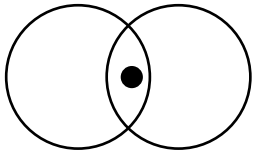
↓

利用变分贝叶斯推断用一个正态分布去**近似**后验分布

↓

$$p(\mathrm{Z}_i|X_i)\sim N(0,1)$$

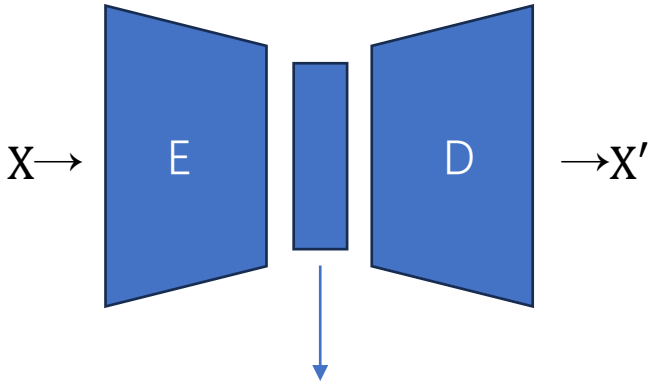
$$\max\ \ loss\ \ = -loss_1 + loss_2 = -KL(q(z|x)||P(z)) + \int q(z|x)\log P(x|z)dz$$



用正态分布去近似合理吗？

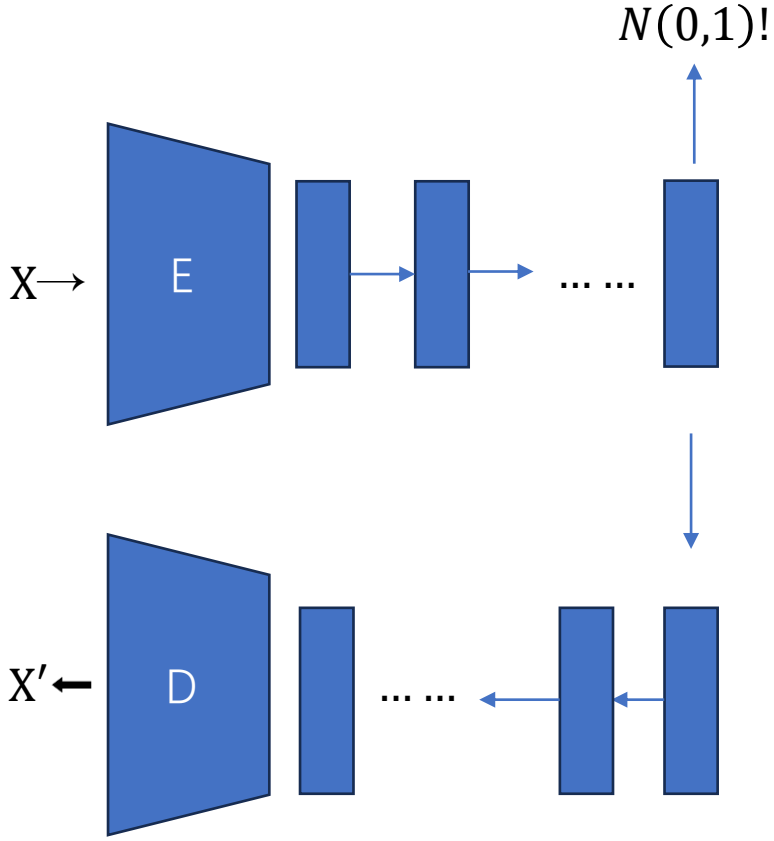
[Auto-Encoding Variational Bayes (arxiv.org)]
ICLR 2014

Review——AE，VAE



$N(0,1)$?

但在推断时是从$N(0,1)$进行采样



Motivation

Does this prompt-based control satisfy our needs?

Can we enable finer grained spatial control by letting users provide additional images that directly specify their desired image composition?
(e.g., edge maps, human pose skeletons, segmentation maps, depth, normals, etc.)

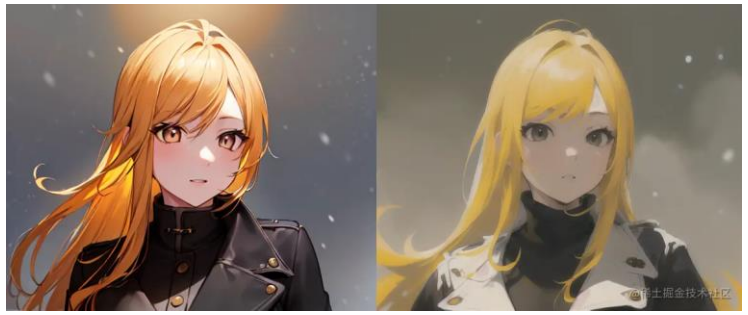
Image-to-image translation models √
A wider variety of problems like depth-to-image, pose-to-image, etc ×

The largest datasets for various specific problems (e.g., object shape/normal, human pose extraction, etc.) are usually about 100K in size, which is 50,000 times smaller than the LAION-5B dataset that was used to train Stable Diffusion.

ControlNet can control Stable Diffusion with various conditioning inputs, like depth-to-image conditioning, training ControlNets on a single NVIDIA RTX 3090Ti GPU can achieve results competitive with industrial models trained on large computation clusters.

Motivation

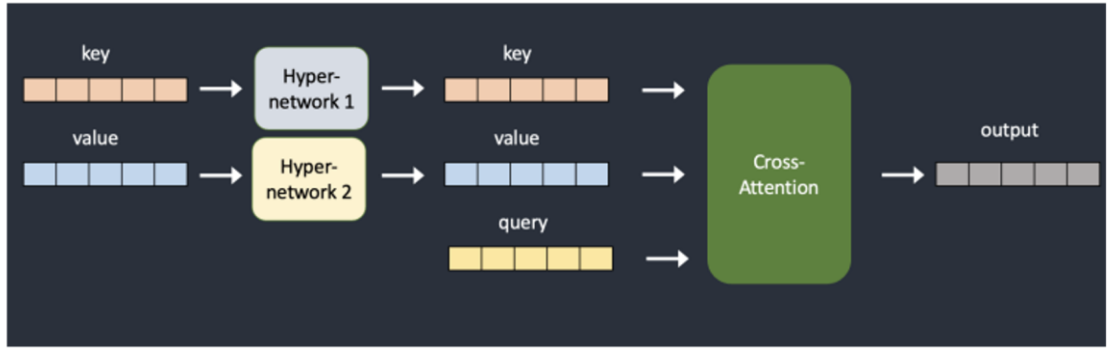
A girl, long hair, blond hair, black eyes, black coat, winter, snow。



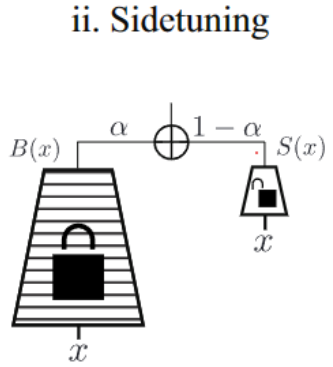
A girl, long hair, black eyes

Related Work

HyperNetwork



Sidetuning



Adapter

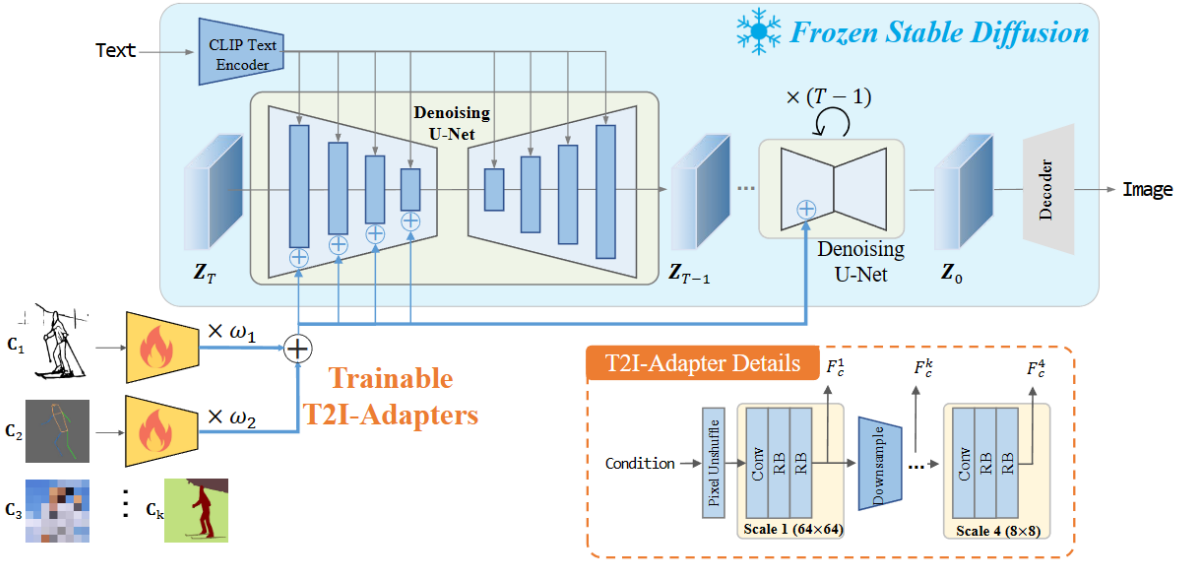
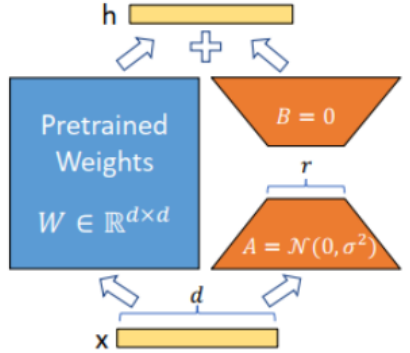


Figure 3. The overall architecture is composed of two parts: 1) a pre-trained stable diffusion model with fixed parameters; 2) several T2I-Adapters trained to align internal knowledge in T2I models and external control signals. Different adapters can be composed by directly adding with adjustable weight ω. The detailed architecture of T2I-Adapter is shown in the lower right corner.

LoRA



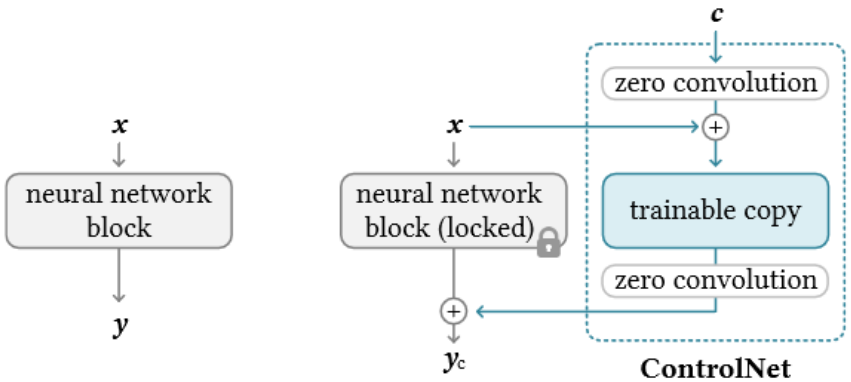
受instrisic dimension工作的启发，作者认为参数更新过程中也存在一个‘内在秩’。对于预训练权重矩阵$W_0 \in \mathbb{R}^{d\times k}$，我们可以用一个低秩分解来表示参数更新$\Delta W$，即：

$$W_0 + \Delta W = W_0 + BA \qquad B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k} \quad and \quad r \ll min(d,k) \quad (3)$$

训练过程中冻结参数W_0，仅训练A和B中的参数。如上图所示，对于$h = W_0 x$，前向传播过程变为：

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad (4)$$

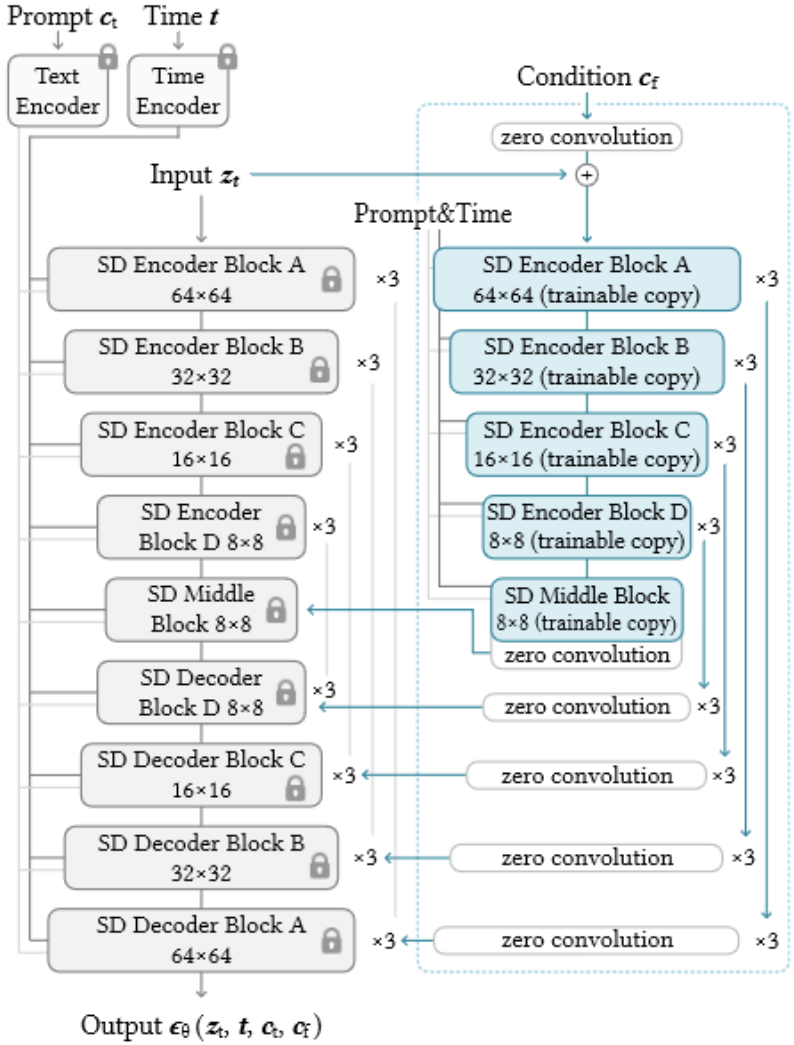
Method



(a) Before (b) After

Figure 2: A neural block takes a feature map x as input and outputs another feature map y, as shown in (a). To add a ControlNet to such a block we lock the original block and create a trainable copy and connect them together using zero convolution layers, *i.e.*, 1×1 convolution with both weight and bias initialized to zero. Here c is a conditioning vector that we wish to add to the network, as shown in (b).

Apply to SD



(a) Stable Diffusion (b) ControlNet

Figure 3: Stable Diffusion's U-net architecture connected with a ControlNet on the encoder blocks and middle block. The locked, gray blocks show the structure of Stable Diffusion V1.5 (or V2.1, as they use the same U-net architecture). The trainable blue blocks and the white zero convolution layers are added to build a ControlNet.

Method

In the training process, we randomly replace 50% text prompts ct with empty strings.

We observe that the model does not gradually learn the control conditions but abruptly succeeds in following the input conditioning image; usually in less than 10K optimization steps. As shown in Figure 4, we call this the “sudden convergence phenomenon”.

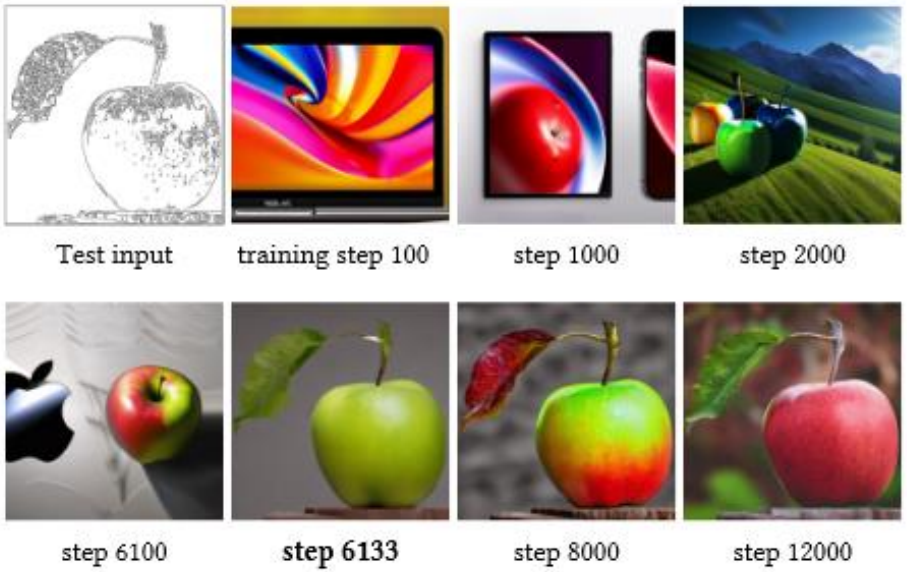


Figure 4: The sudden convergence phenomenon. Due to the zero convolutions, ControlNet always predicts high-quality images during the entire training. At a certain step in the training process (*e.g.*, the 6133 steps marked in bold), the model suddenly learns to follow the input condition.

Method

Composing multiple ControlNets

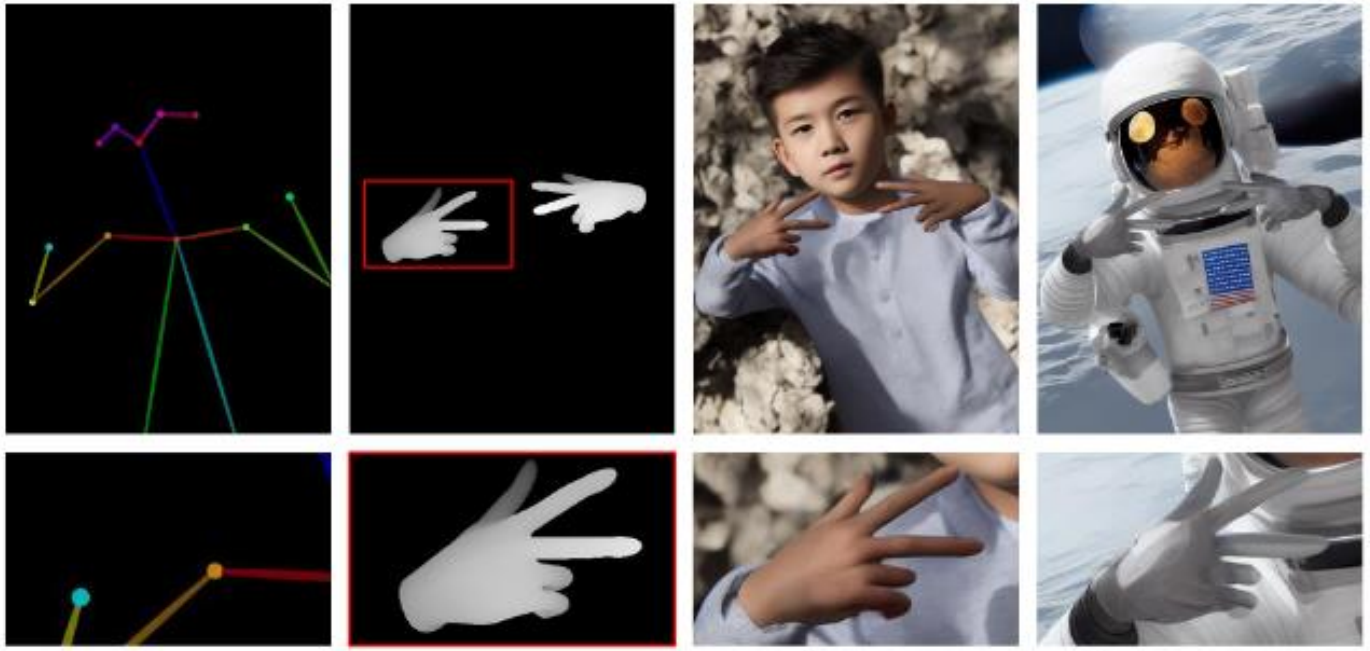


Figure 6: Composition of multiple conditions. We present the application to use depth and pose simultaneously.

Experiments —— Qualitative Result

Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc.*, to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

Experiments —— Qualitative Result

Figure 7: Controlling Stable Diffusion with various conditions **without prompts**. The top row is input conditions, while all other rows are outputs. We use the empty string as input prompts. All models are trained with general-domain data. The model has to recognize semantic contents in the input condition images to generate images.

Experiments —— Ablative Study

Figure 8: Ablative study of different architectures on a sketch condition and different prompt settings. For each setting, we show a random batch of 6 samples without cherry-picking. Images are at 512 × 512 and best viewed when zoomed in. The green "conv" blocks on the left are standard convolution layers initialized with Gaussian weights.

Experiments —— Discussion

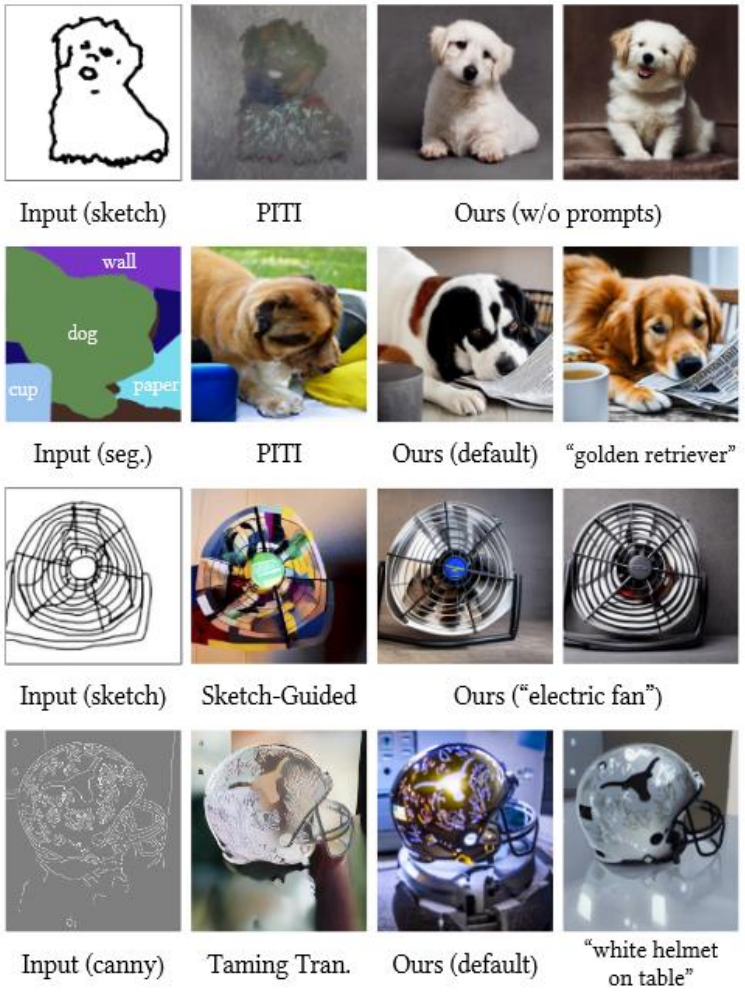


Figure 9: Comparison to previous methods. We present the qualitative comparisons to PITI [89], Sketch-Guided Diffusion [88], and Taming Transformers [19].

- **Dataset sizes:** The training does not collapse with limited 1k images. The learning is scalable when more data is provided.
- **Capability to interpret contents**:

Figure 11: Interpreting contents. If the input is ambiguous and the user does not mention object contents in prompts, the results look like the model tries to interpret input shapes.

- **Transfer ability**: Since ControlNets do not change the network topology of pretrained SD models, it can be directly applied to various models in the stable diffusion community.

Related Work

Deep Unsupervised Learning using Nonequilibrium Thermodynamics (arxiv.org) ICML 2015

Generating Images from Captions with Attention (arxiv.org) ICLR 2016

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models (arxiv.org) CVPR 2021

Diffusion Models Beat GANs on Image Synthesis (arxiv.org) NeurIPS 2021

Taming Transformers for High-Resolution Image Synthesis (arxiv.org) (VQ-GAN) CVPR2021

Zero-Shot Text-to-Image Generation (arxiv.org) (DALL·E)

Hierarchical Text-Conditional Image Generation with CLIP Latents (arxiv.org) (DALL·E-2) CVPR2022

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (arxiv.org) (Imagen) NeruIPS2022

使用 🧨 Diffusers 实现 ControlNet 高速推理 (huggingface.co)

西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions

主讲人：XXX

2024. XX. XX