文生图扩散模型中的条件控制

汇报人：李泽信

Adding Conditional Control to Text-to-Image Diffusion Models
向文生图扩散模型中添加条件控制

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University

ICCV2023 Best Paper (Marr Prize)

Diffusion Model 的扩散过程（逆向过程）

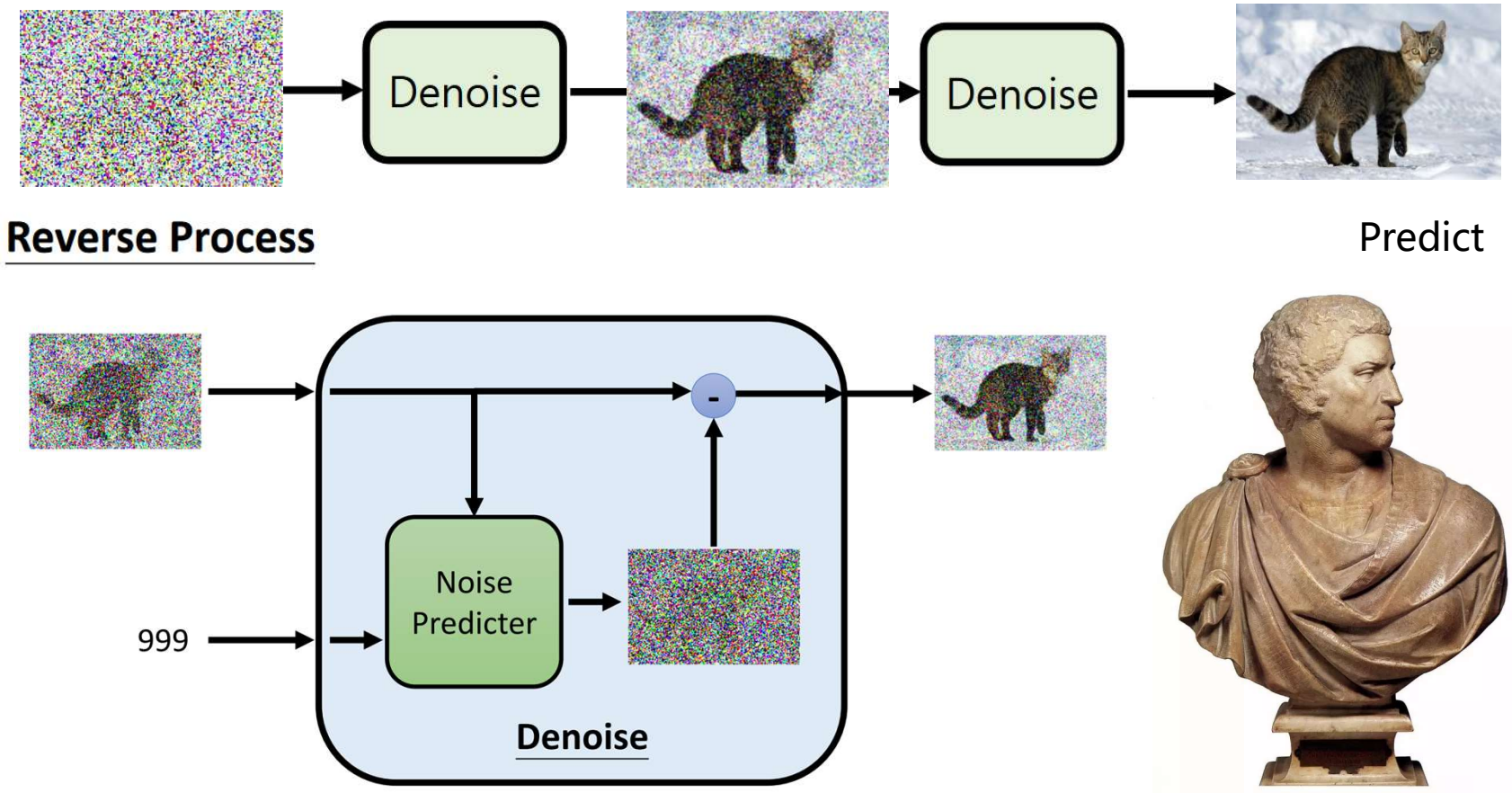


image source: https://speech.ee.ntu.edu.tw/~hylee/ml/ml2023-course-data/DiffusionModel%20(v2).pdf

文生图开源模型
Stable Diffusion 的结构

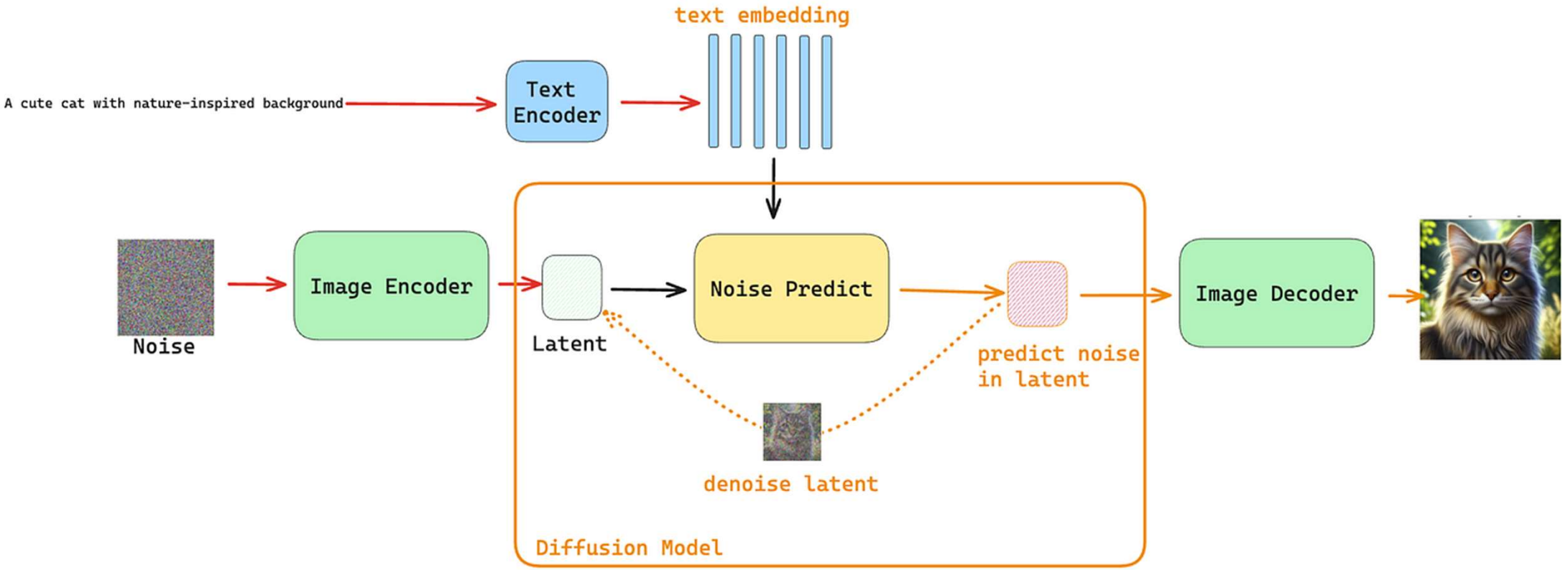


image source: https://medium.com/thedeephub/a-comprehensive-guide-to-diffusion-model-1-ddpm-ee68ff2a6a4b

Stable Diffusion 的 Noise Predictor: U-Net
控制条件是怎么添加进来的呢？

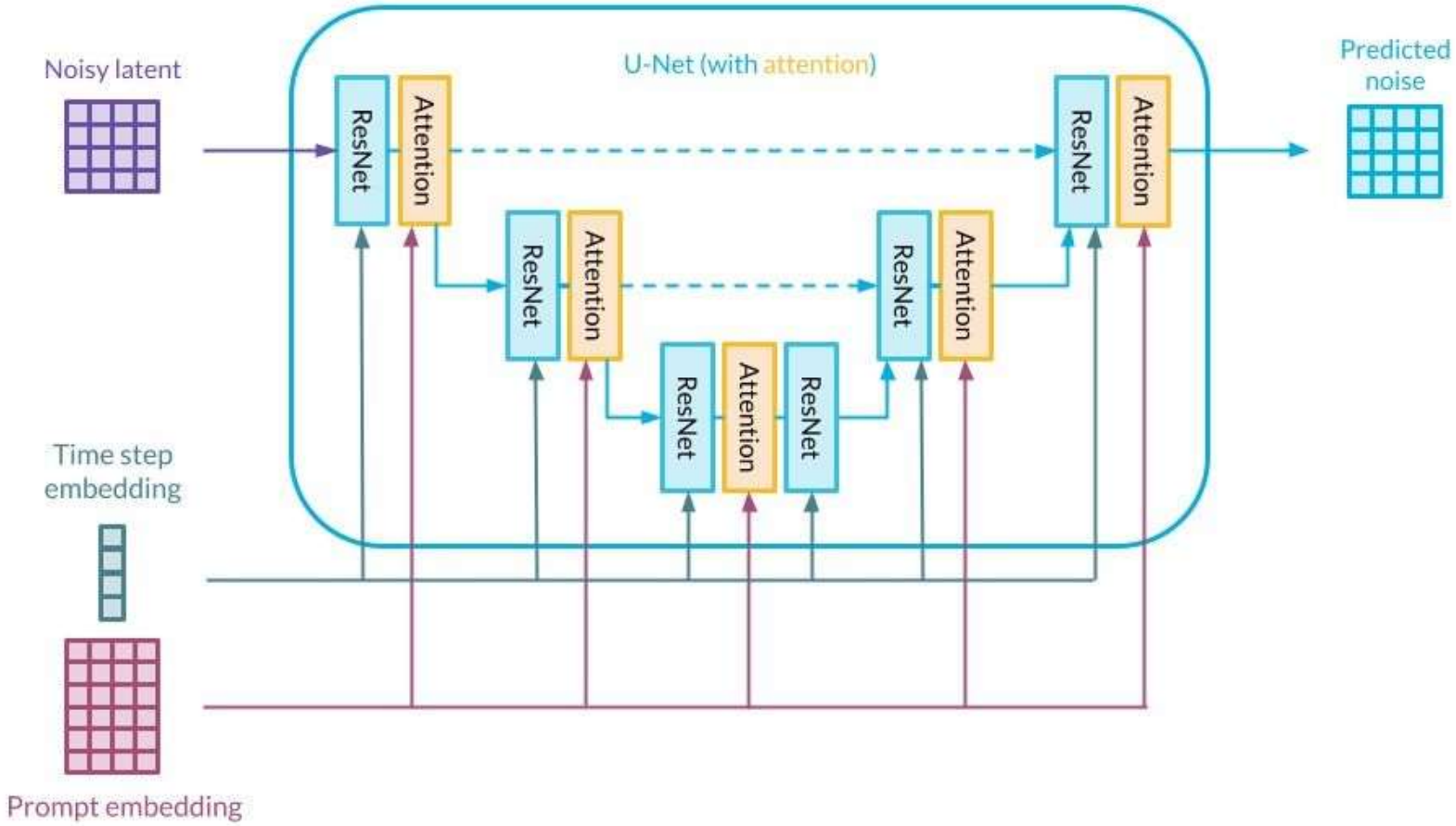


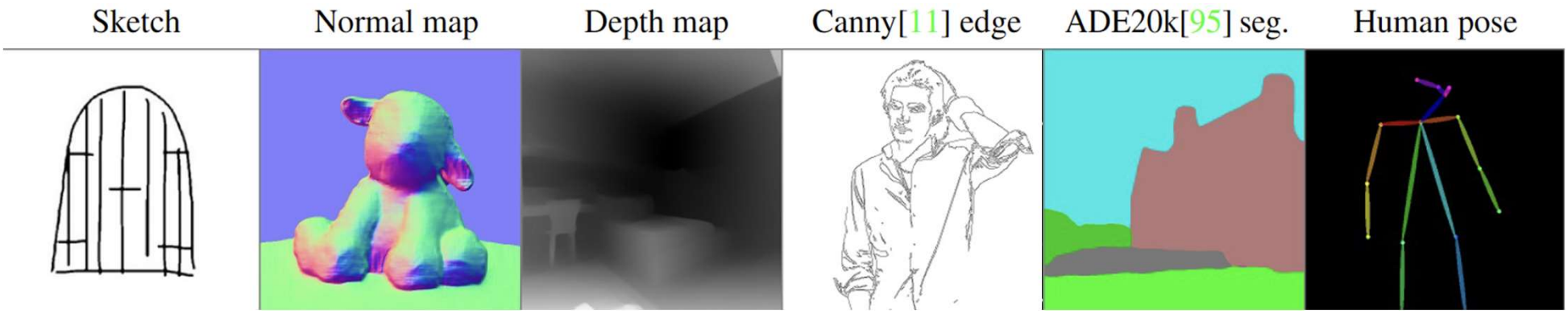
image source: https://deepsense.ai/diffusion-models-in-practice-part-1-the-tools-of-the-trade/

Stable Diffusion 效果展示

image source: yuque/@toulzx

Motivation

- 仅通过文本提示不足以精确表达复杂的布局、姿势、形状和形式。
- **设想：通过提供额外的图片，实现更精细的空间控制。**
  - pose, segmentation, depth, etc.
  - 需要以端到端的方式学习。(end-to-end)



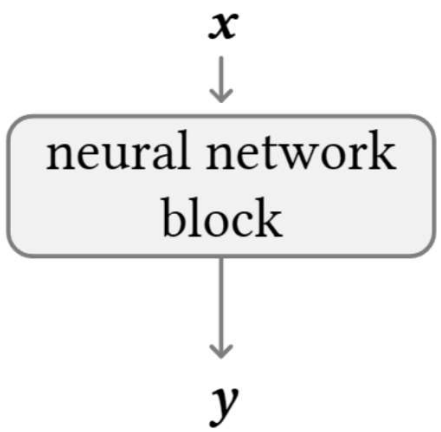
Images are from the original paper if not stated.

Input human pose　Default　“chef in kitchen”　“Lincoln statue”

- 图生图模型，学习从条件图像到目标图像的映射。
- 用有限的数据直接微调或继续训练大型预训练模型，可能会导致过度拟合和灾难性遗忘。
  （规模远小于用以训练 SD 的 LAION-5B 数据集）
- **可能需要额外构建一个端到端神经网络架构** => ControlNet

ControlNet 基本结构

- 术语 network block 定义：用以指代常见的神经网络单元。
  - e.g., resnet block, conv-bn-relu block (Conv+BatchNorm+ReLU), multi-head attention block, transformer block, etc.

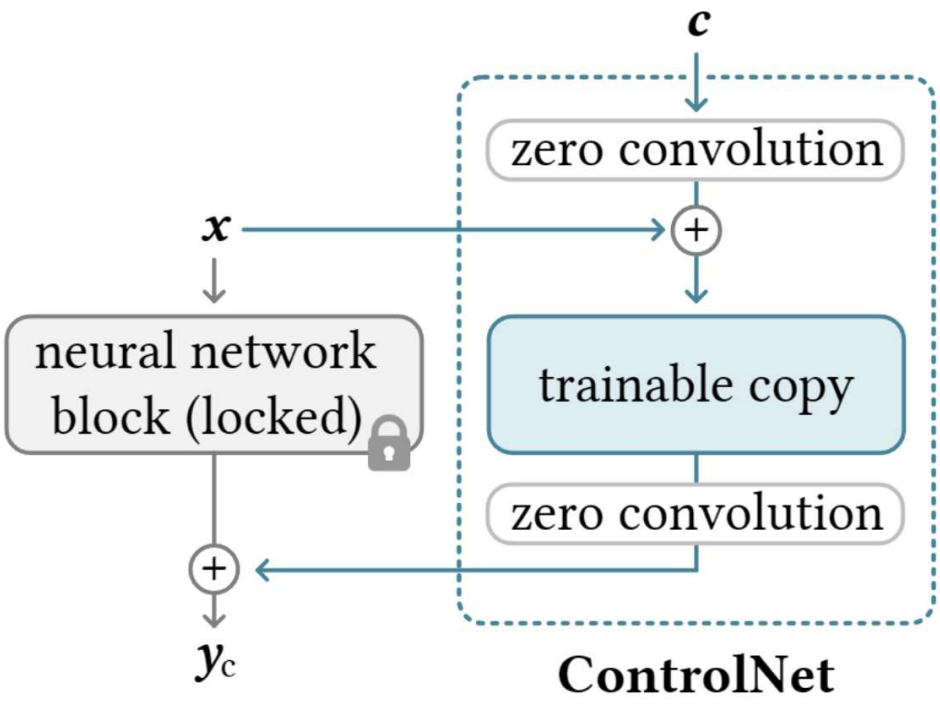


- 假设 $\mathcal{F} = (\cdot;\Theta)$ 是一个训练好的 network block。$\mathcal{F}$ 将输入的 feature map x 转换为另一个 feature map y 可记为：

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x};\Theta)$$

ControlNet 基本结构

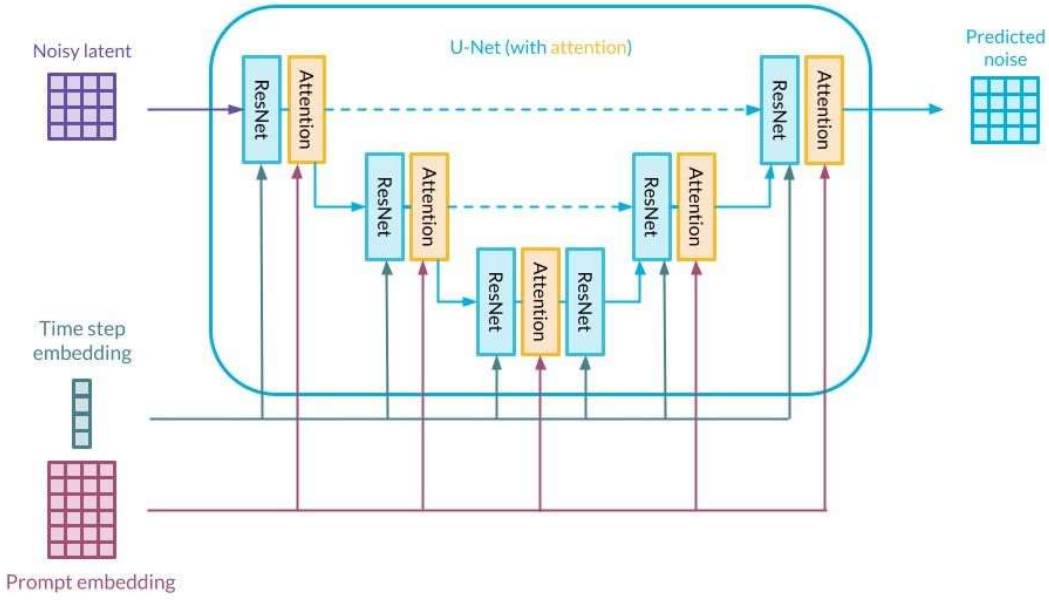
- 冻结 network block（称为原始块）并创建可训练副本（称为 copy 块），再使用零卷积层将它们连接在一起。
- 术语 zero convolution layer 定义：1×1 卷积层,weight 和 bias 均初始化为零。称零卷积层。
- 假设 $Z(\cdot;\Theta_z)$ 为零卷积层，输入 c 是我们希望添加到网络中的条件向量，完整的 ControlNet 会计算并输出：

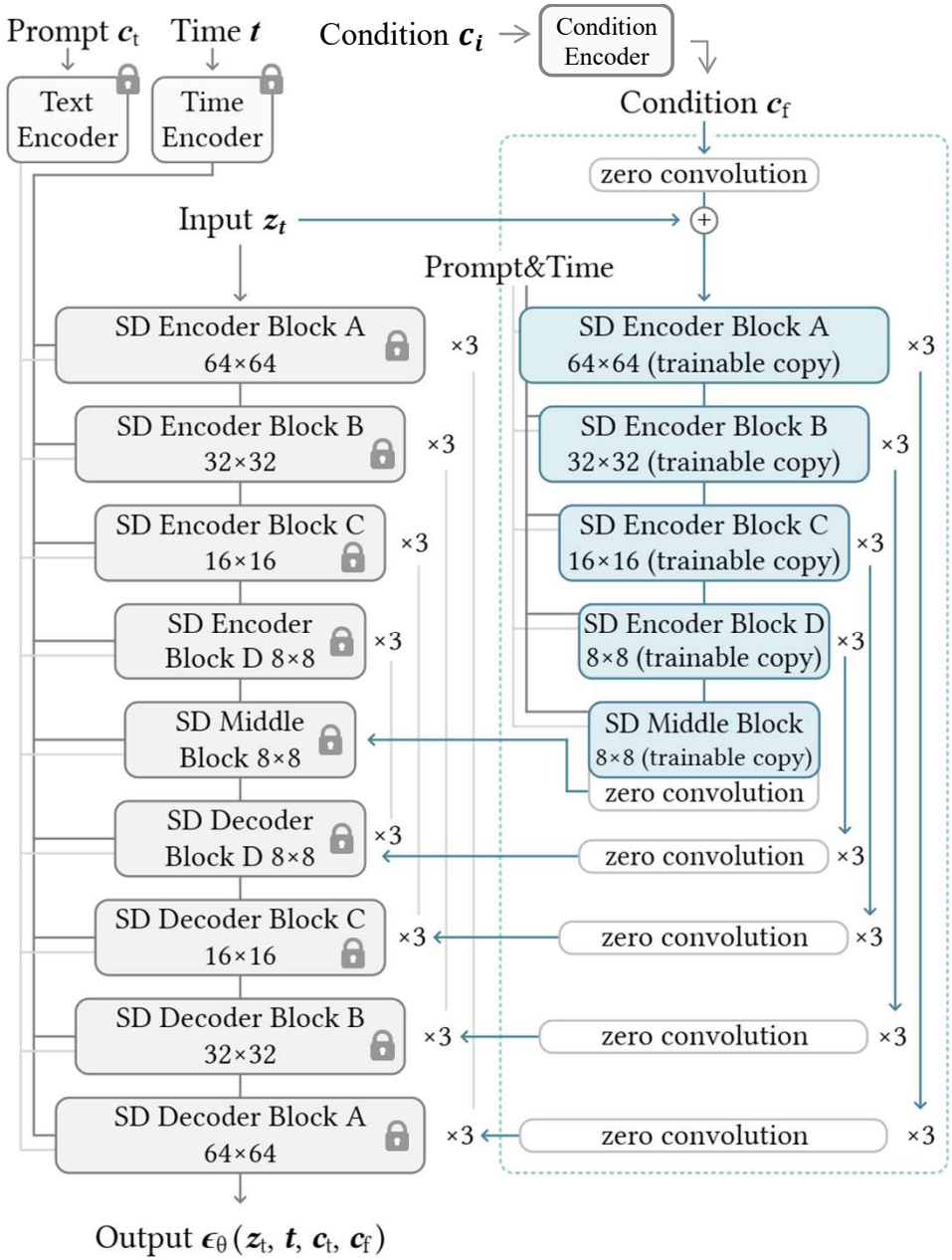


$$\boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x};\Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c};\Theta_{z1});\Theta_c);\Theta_{z2})$$

文生图扩散模型中的 ControlNet

以 Stable Diffusion 为例，
也适用于所有采用 U-net 架构的网络。





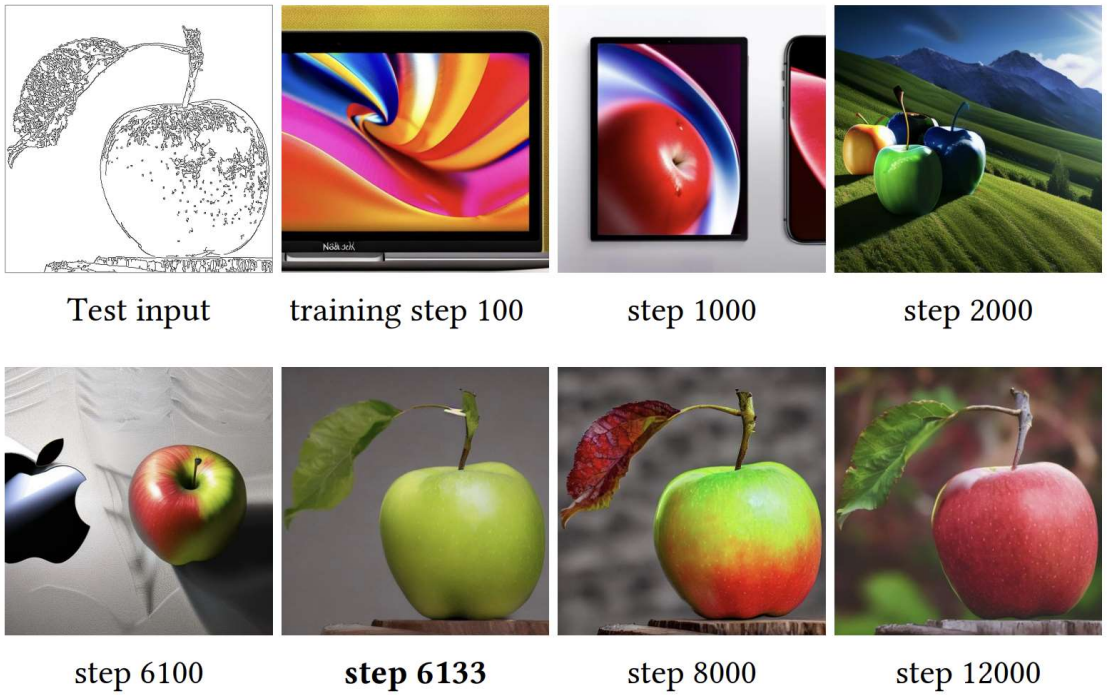
模型训练

- 损失函数： $\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f})) \|^2 \right]$

- **训练技巧**：随机将 50% 的 text prompt c_t 替换为空字符串。这种方法增强了 ControlNet 对 conditioning image 中的语义（例如 edge、pose、depth 等）的识别能力，获得替代 text prompt 的能力。

突然收敛现象

模型并不是逐渐学习 control condition 的，而是突然学到的。
通常 steps<10K。扩大 batch size 重新训练，可获得更好效果。



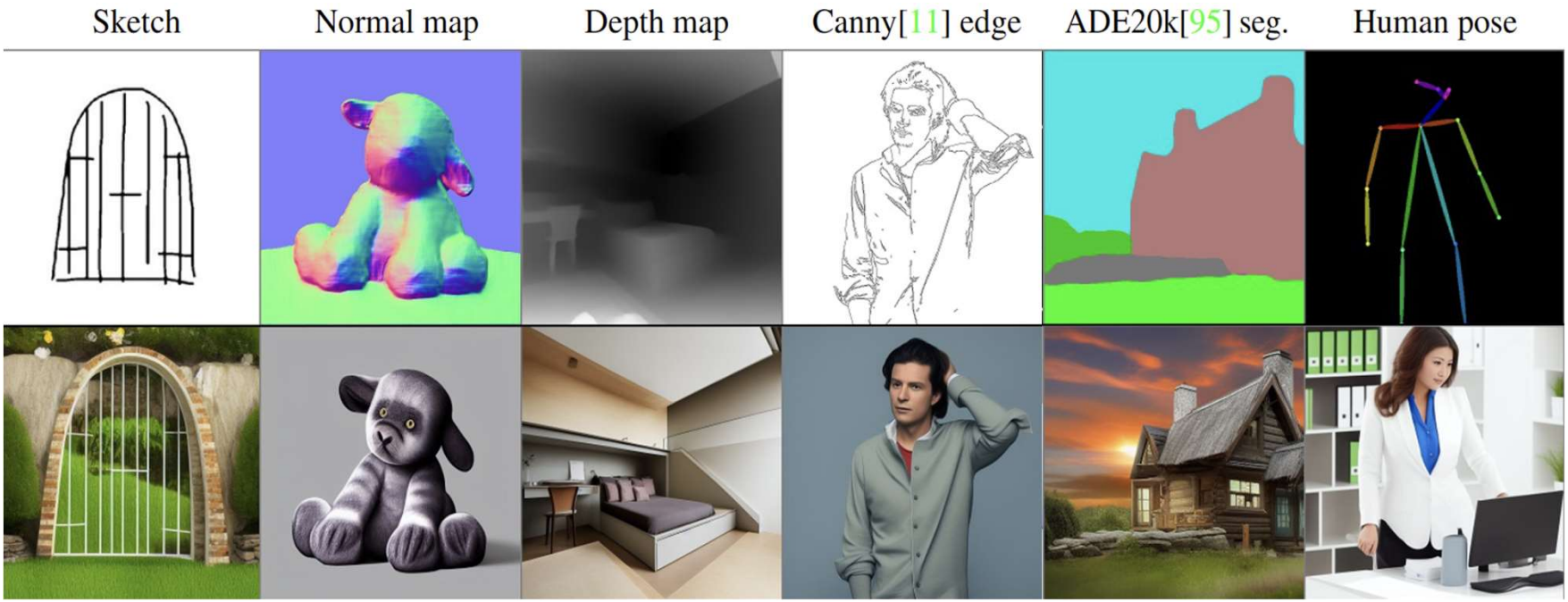
在不同规模数据集上训练的建议

- Small-Scale Training：部分断开 ControlNet 和 Stable Diffusion 之间的连接可以加速收敛。当模型显示出结果与条件之间的合理关联时，就可以在继续训练中再次连接这些断开的链接，以促进精确控制。
- Large-Scale Training：可以先对 ControlNet 进行足够多的迭代训练（通常超过 50k 步），然后解锁 Stable Diffusion 的所有权重，并将整个模型作为一个整体进行联合训练。这将会产生一个更加针对具体问题的模型。

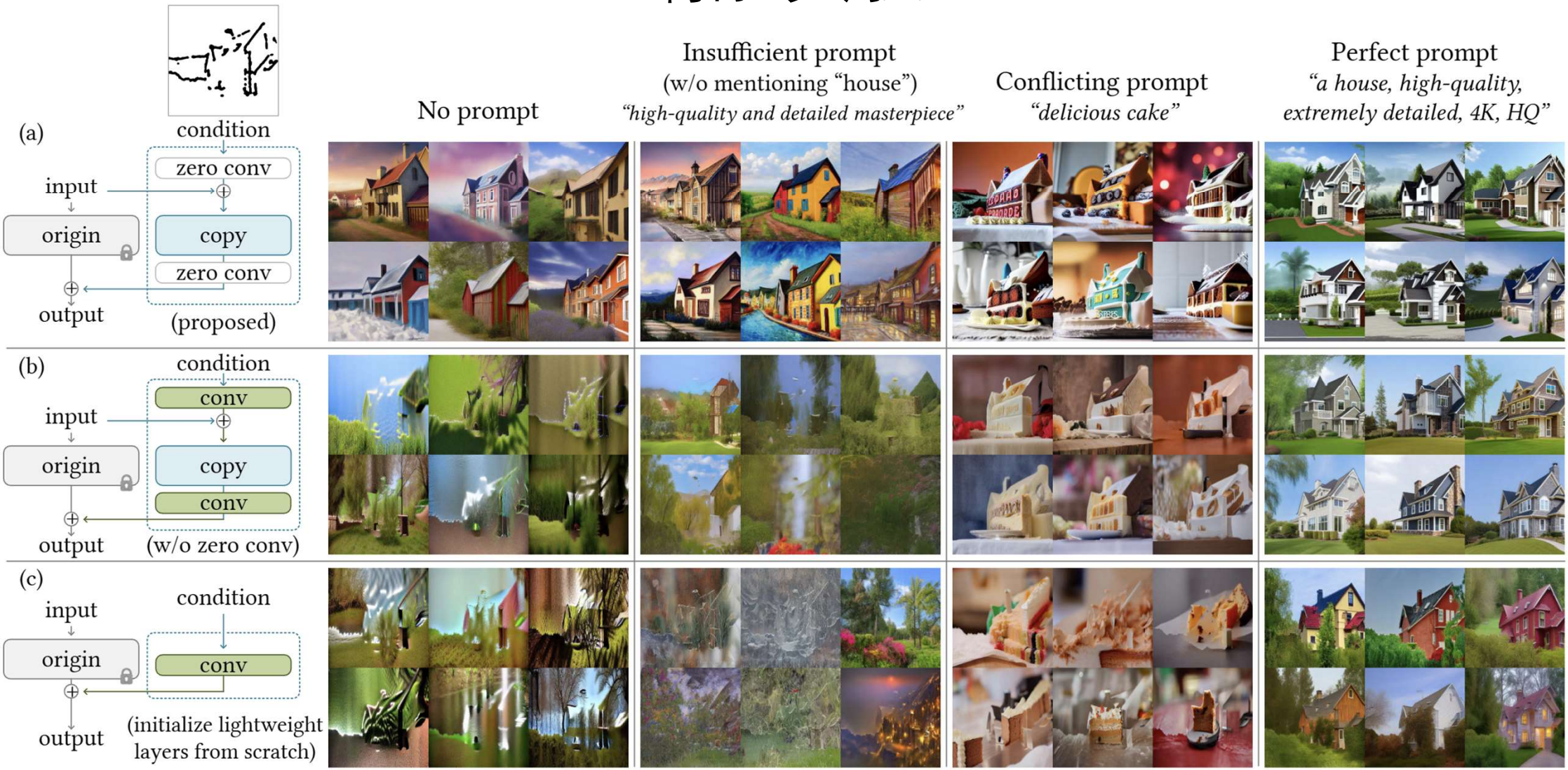
实验定性结论：ControlNet 学到了内容语义

没有 text prompt 时，各种条件下的结果。

其中 ControlNet 解释了不同输入的 conditioning image 中的内容语义。



消融实验


No prompt
Insufficient prompt
(w/o mentioning "house")
"high-quality and detailed masterpiece"
Conflicting prompt
"delicious cake"
Perfect prompt
"a house, high-quality, extremely detailed, 4K, HQ"
(a)
condition
zero conv
input
origin
copy
zero conv
output
(proposed)
(b)
condition
conv
input
origin
copy
conv
output
(w/o zero conv)
(c)
input
condition
origin
conv
output
(initialize lightweight layers from scratch)

消融实验结论

- (c) 仅使用普通卷积层作为 ControlNet 结构训练，模型不足以 interpret 输入的 conditioning image，并且在不充分 prompt(ii)和无 prompt(i)的情形下失败。
  作者称这种架构为 ControlNet-Lite。

- (b) 使用普通卷积层替换零卷积层，模型性能下降，不足以 interpret 输入的 conditioning image，这说明在 finetuning 过程中破坏了预训练好的 copy 块的原有能力。

量化评价：比肩工业模型效果 (Depth)

- Stable Diffusion V2 Depth-to-Image (SDv2-D2I) 模型，
  在 NVIDIA A100 集群上训练上千个 GPU hours，
  训练数据>12M。
- ControlNet for SDv2 with depth conditioning 模型，
  在单张 NVIDIA RTX 3090Ti 上训练 5 天，
  训练数据为 200K。

各生成 100 张图，让 12 名用户学习所属关系；

然后再各生成 200 张图，让用户判断图片与模型的对应关系。

最终平均准确率是 0.52 ± 0.17，

这说明两个模型生成的图片效果是几乎一致的（但是 ControlNet 却拥有更小的训练成本）。

量化指标:

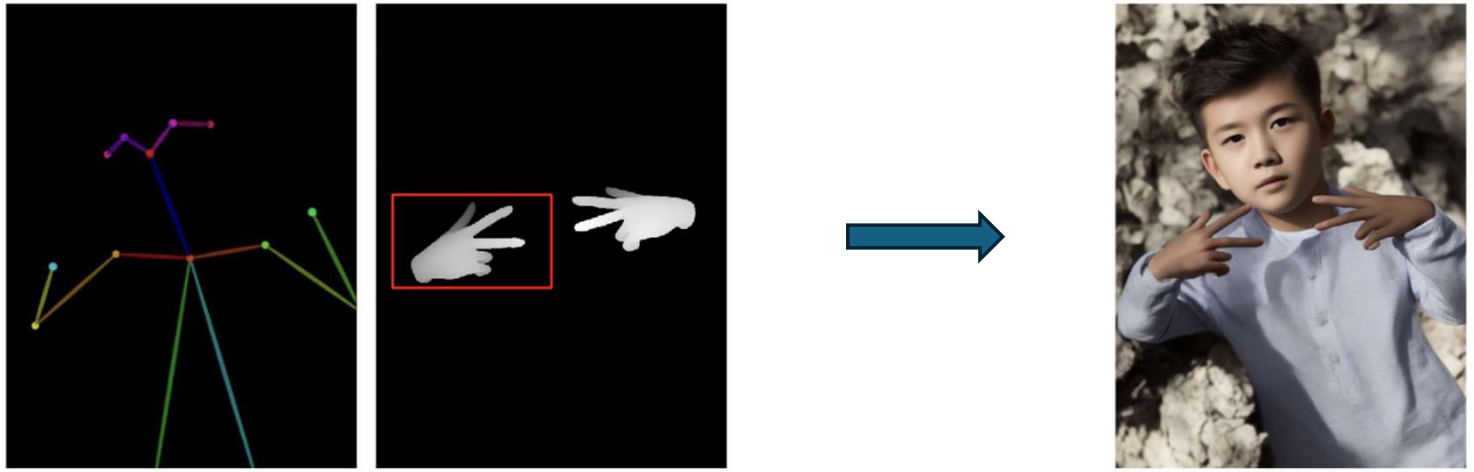
FID score、CLIP score、CLIP aesthetic score (Segment)

Method	FID ↓	CLIP-score ↑	CLIP-aes. ↑
Stable Diffusion	6.09	0.26	6.32
VQGAN [19](seg.)*	26.28	0.17	5.14
LDM [71](seg.)*	25.35	0.18	5.15
PIPT [88](seg.)	19.74	0.20	5.77
ControlNet-lite	17.92	0.26	6.30
ControlNet	15.27	0.26	6.31

- FID：计算生成图集和真实图集之间的分布距离。
- CLIP 是用 400M image-text Pair 训练出来的模型，CLIP score 评价的是通过 text encoder 和 image encoder 后两向量的相似程度。
- AESTHETIC_SCORE：laion5b 数据集中的图片的美学评分。首先选择一小部分图片数据集让人对图片的美学打分，然后基于这个标注数据集来训练一个打分模型，并对所有样本计算估计的美学评分。

ControlNet 支持组合

将多个 conditioning image（例如 pose 和 depth）应用于 Stable Diffusion 的单个实例，我们可以直接将相应 ControlNet 的输出添加到 Stable Diffusion。这种组合不需要额外的加权或线性插值。



ControlNet 在不同规模数据集上具有鲁棒性

- 在只有 1k 张图像的情况下，训练并不会崩溃，而且还能让模型生成可识别的内容。
- 当提供更多数据时，学习是可扩展的。

“Lion” 1k images 50k images 3m images

ControlNet 可直接移植到社区模型上

- ControlNet **不改变**预训练 Stable Diffusion 模型的网络拓扑结构。
- C 站那些基于 SD 微调的模型都可以**即插即用。**
- 事实上 ControlNet 已经成为“ai 作图游戏规则的改变者”

总结

- ControlNet 是一种神经网络架构，可实现在大型预训练文生图模型上的条件控制。
- 它重用原始模型的大规模预训练层，利用经过数十亿幅图像预训练的编码层，作为学习各种条件控制的基础。
  它通过使用零卷积层连接，确保微调过程不受噪音干扰。
- 大量实验表明，ControlNet 能捕获语义、适应不同规模的数据集、在多种任务上表现出色，媲美特定任务的工业模型。
- 即插即用，ControlNet 可直接移植应用于大量社区模型上。

ControlNeXt: Powerful and Efficient Control for Image and Video Generation
强大而高效的图像和视频生成控制

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang
Jiaya Jia (贾佳亚)
The Chinese University of Hong Kong

ArXiv v2 *(Aug.15, 2024)*

思考 1：再谈“突然收敛”...

- 模型最初无法学习控制能力，然后突然获得这种技能。

Images are from the original paper if not stated.

“突然收敛”现象的本质

- 直接组合随机初始化得到的新参数，通常会导致训练崩溃和收敛性差，原因是引入的模块和预训练模型之间的**数据分布不一致**。
- 零初始化确保了训练开始时不会受到新引入的模块的影响，但仍会导致收敛缓慢，具体表现为“突然收敛”。

1. 零卷积抑制了损失函数的影响，导致 warm-up 阶段延长，模型难以开始有效学习。
2. 预训练的生成模型完全冻结，ControlNet 仅充当 Adapter，不会立即影响模型。

“交叉归一化”替代“零卷积”

- 归一化方法 （如 batch ~ 和 layer ~）标准化层输入以提高训练稳定性和速度。它们通过将输入归一化为零均值和单位方差来实现这一点，这在神经网络训练中被广泛使用。

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

- 受此启发，提出交叉归一化来对齐处理后的条件控制和主分支特征，确保训练稳定性和速度。

$$\hat{\boldsymbol{x}}_c = \frac{\boldsymbol{x}_c - \boldsymbol{\mu}_m}{\sqrt{\boldsymbol{\sigma}_m^2 + \epsilon}} * \gamma, \qquad (10)$$

Above Algorithm comes from the paper of Batch Normalization: https://arxiv.org/pdf/1502.03167

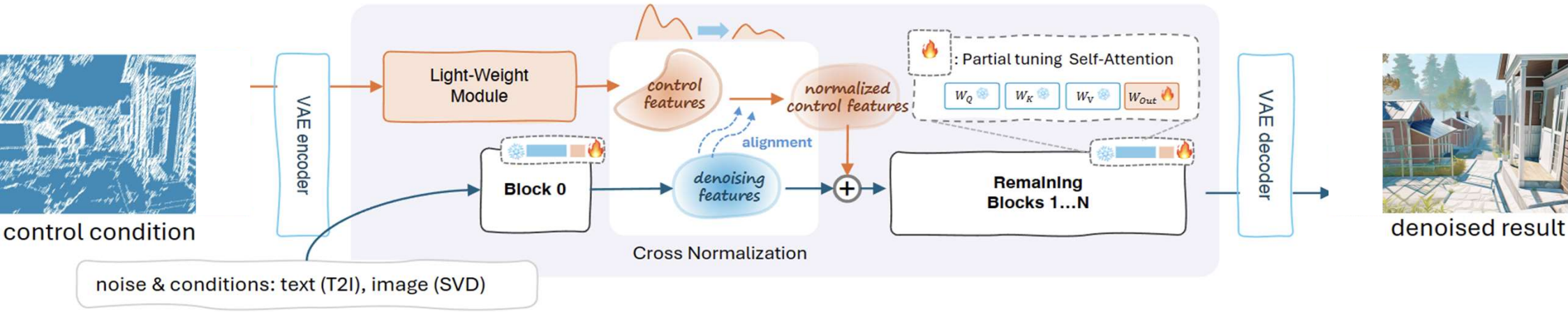
思考 2：Copy 块是必须的吗？

- 预训练的大型生成模型已足够强大，无需引入如此大量的额外参数来实现控制生成能力，依赖生成模型本身来处理控制信号即可。
- 用仅由多个 ResNet 块组成的轻量级卷积模块替换 Copy 块。
- 对比 ControlNet 消融实验...

Pipeline of ControlNeXt

- 控制通常具有简单的形式，或与去噪特征保持高度的一致性，从而无需在多个阶段插入控制。

- 对比 ControlNet 结构...



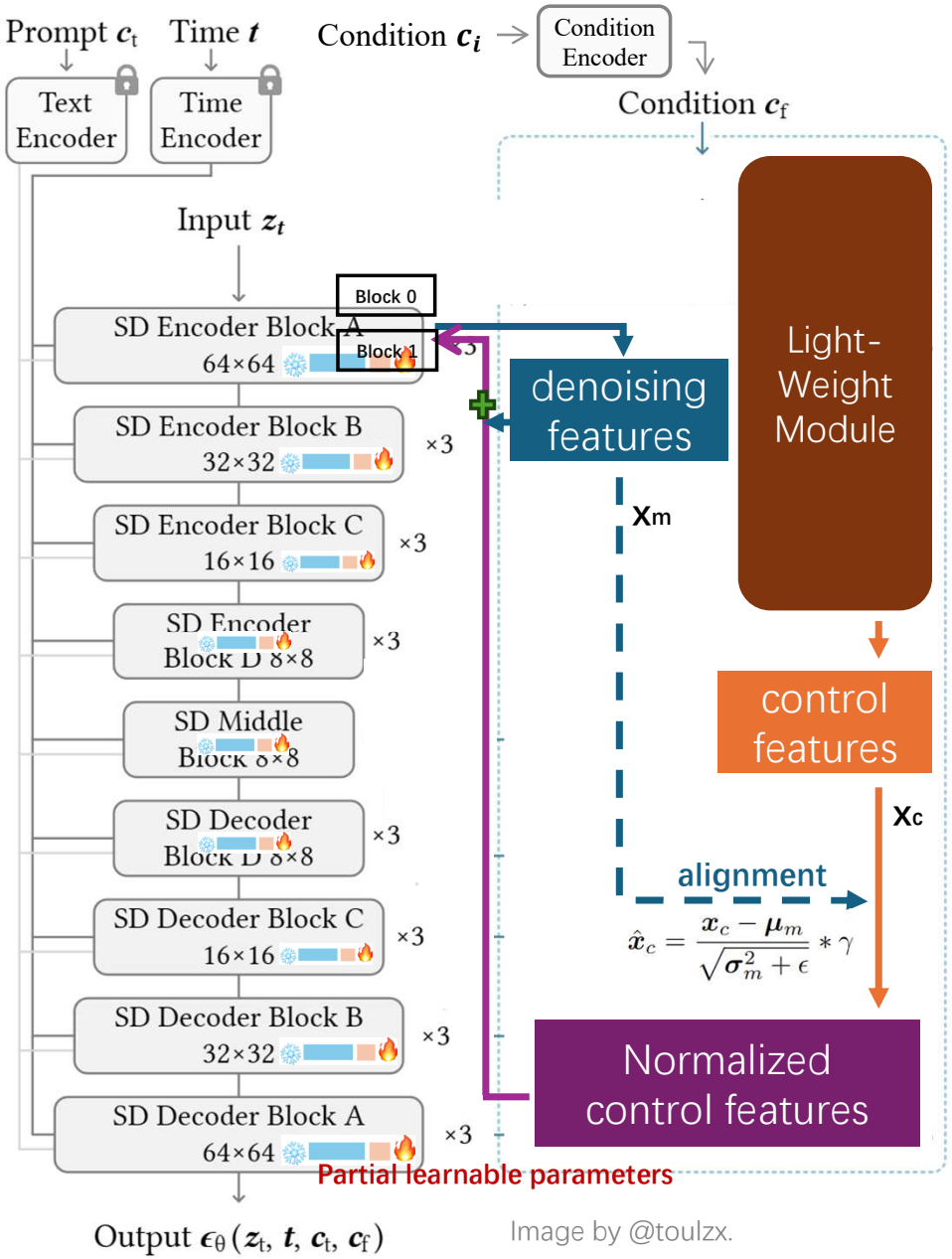
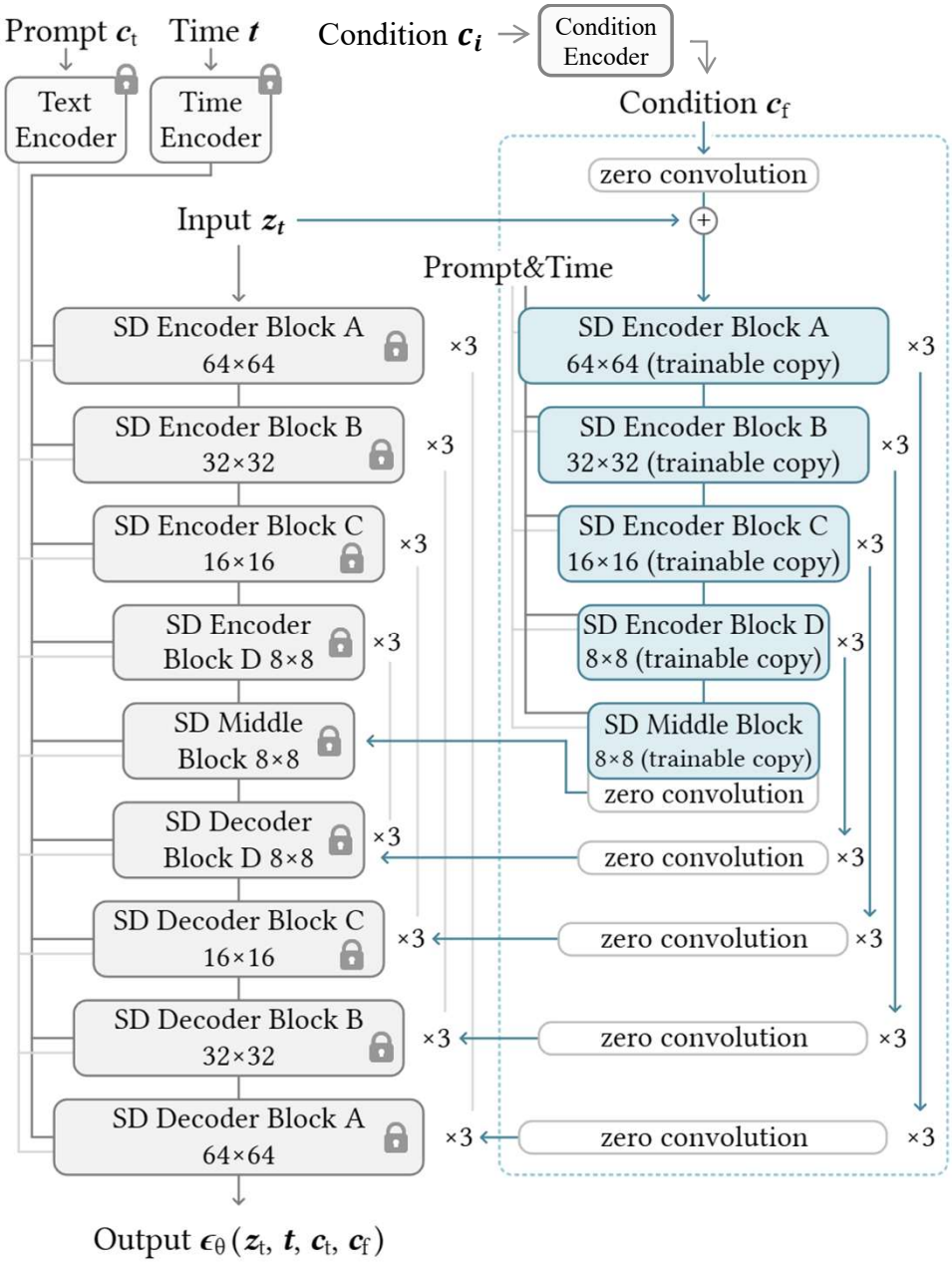


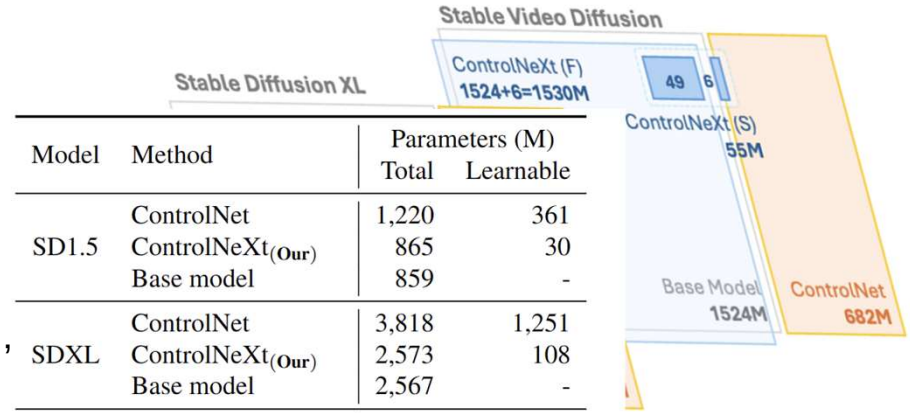
Image by @toulzx.

V.S.



提升

- 最多需要不到 10% 的可学习参数。
- 经过训练的 base model，即插即用？
  - 尽管基础模型的一小部分参数子集需要参与训练，但是在单个主干上训练得到的 ControlNeXt 在不同主干仍保持了较好的控制能力。



Model	Method	Parameters (M) Total	Learnable
SD1.5	ControlNet	1,220	361
	ControlNeXt$_{(Our)}$	865	30
	Base model	859	-
SDXL	ControlNet	3,818	1,251
	ControlNeXt$_{(Our)}$	2,573	108
	Base model	2,567	-
SVD	ControlNet	2,206	682
	ControlNeXt-S$_{(Our)}$	1,530	55
	ControlNeXt-F$_{(Our)}$	1,530	1,530
	Base model	1,524	-

- 附加的控制模块时间开销：40%=>10%（对视频生成这很重要！帧）

Method	Inference Time (s) SD1.5	SDXL	SVD	Δ
ControlNet	0.31	1.01	1.73	+ 41.9%
ControlNeXt$_{(Our)}$	0.24	0.82	1.29	+ 10.4%
Base model	0.22	0.70	1.23	-

买家秀和卖家秀

Project Page: https://pbihao.github.io/projects/controlnext/index.html
Videos from #issue1: https://github.com/dvlab-research/ControlNeXt/issues/1

“Currently, our work does not specifically address IP consistency.”

原始照片

买家秀

买家秀

总结

- ControlNeXt 是一种**更轻量的**神经网络架构，可实现在大型预训练文生图模型上的条件控制。
- 它**直接利用**大规模预训练模型，实现各种条件控制，通过轻量的神经网络提取控制特、通过**交叉归一化**对齐特征，确保训练稳定性和速度。
- 即插即用，ControlNeXt 可直接应用于大量社区模型上。

Something else...

大模型的摩尔定律：Scaling Law

- 2020年，OpenAI在一篇论文中提出一个定律：Scaling law。
  这个定律指的是大模型的最终性能主要与计算量、模型参数量和训练数据量三者的大小相关，而与模型的具体结构（层数/深度/宽度）基本无关。
- 具体来说，当不受其他因素制约时，模型的性能与这三者呈现幂律关系。这意味着，增加计算量、模型参数量或数据大小都可能会提升模型的性能，但是提升的效果会随着这些因素的增加而递减。
- 这说明：随着时间推移，当模型参数规模达到一定程度时，性能提升速度可能会放缓。同时高质量训练数据的持续获取也是亟需解决的一大难题。

- AI 行业正经历模型规模下行的压力，过去一年大部分开发工作落在了小模型上，比如 Anthropic 的 Claude 3.5 Sonnet、Google 的 Gemini 1.5 Pro， OpenAI 的 GPT-4o mini。

For More：https://blog.csdn.net/qq_32907491/article/details/138449138

贾佳亚团队：少参数、小算力、大成果

- 2023 年 10 月，LongLoRA （ICLR 2024 Oral）：
  在单台 8x A100 设备上，LongLoRA 将 LLaMA2 7B 从 4k 上下文扩展到 100k， LLaMA2 70B 扩展到 32k。
  （实现用Llama2-13B总结《三体》）
- 2023 年 12 月， LLaMA-VID（ECCV 2024）：
  处理长视频因视觉 token 过多导致的计算负担（每帧压缩至2个）。
  （由单个 3090 GPU 实现的 Demo，支持 30 分钟的视频处理）
- 2024 年 4 月，Mini-Gemini：
  仅使用 2-3M 数据便实现了对图像理解、推理和生成的统一流程。
  （*人称“小 GPT-4 + DALL-E3”，即将支持语音*）

谢谢