OpenAI

SORA

视频生成原理剖析

ZOMI

SORA Talk Overview

1. **官网解读**：SORA 效果预览 & 技术报告解读
2. **技术架构**： SORA 关键技术方向 & 训练流程
3. **一些思考**：算力规模增长 & 市场预测

SORA 解读

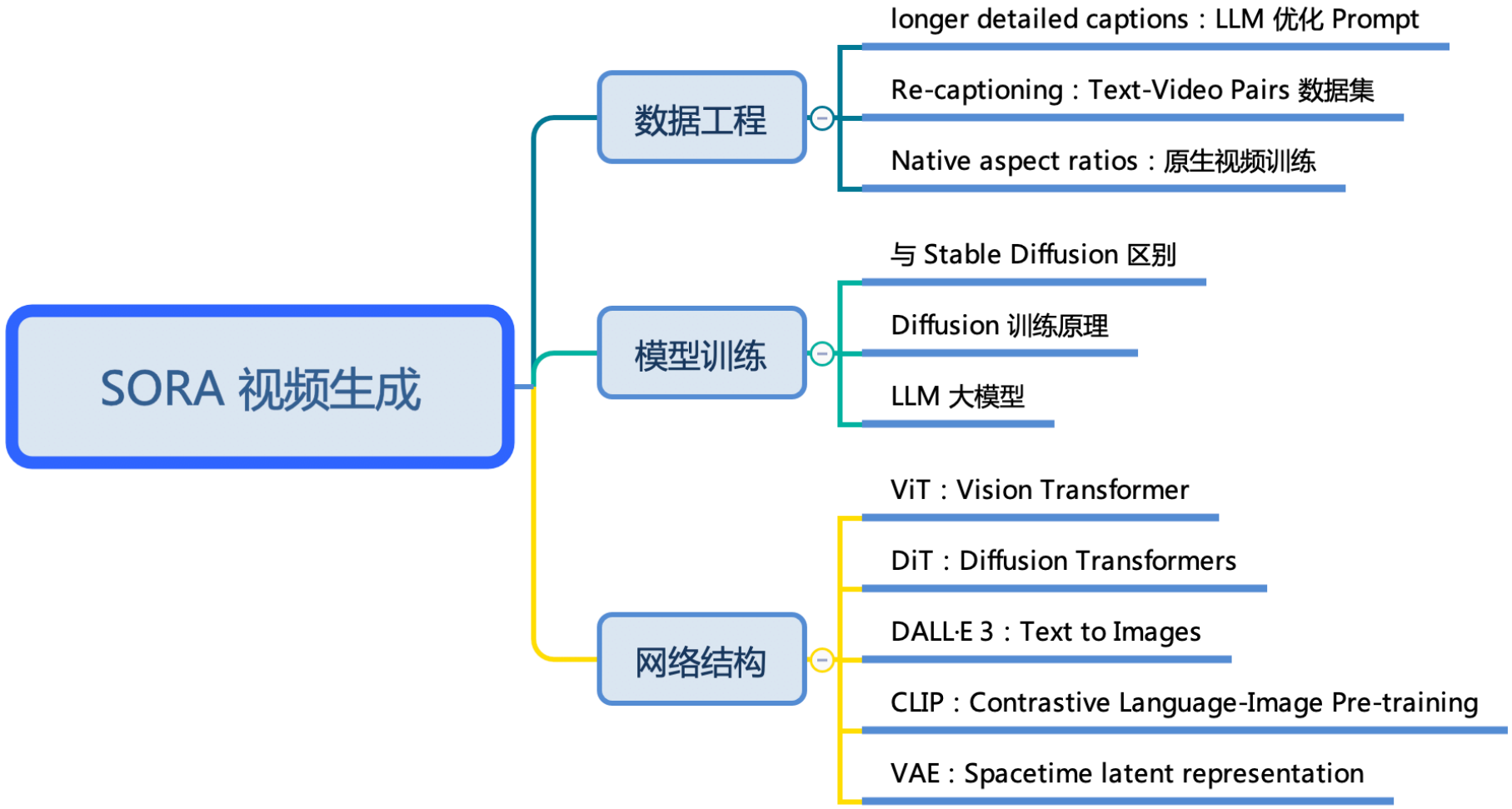
省流总结

https://openai.com/research/video-generation-models-as-world-simulators

1. **内容上：**
   - 最大支持60秒高保真视频生成，支持短视频前后扩展，即可保持视频连续，并扩展时长；
   - 支持基于视频 +文本视频编辑，一句话改变原视频，彻底改变视频创作方式；
2. **技术上：**
   - 将视频压缩为空间时间块（Spacetime patches），使用 Diffusion Transformer 作为主干网络建模；
   - 由于将视频信息压缩为 lower-dimensional latent space，可支持不同尺寸、时间、分辨率的直接生成；
3. **数据工程：**
   - 使用 DALL·E 3 进行视频文本标注；
   - 利用 GPT4 将用户输入的简短提示词，扩充为复杂细节文本；
4. **其他：**
   - 物理交互的细节仍有缺陷，如玻璃破碎与水流，雪地脚印无法生成等；

SORA 技术架构

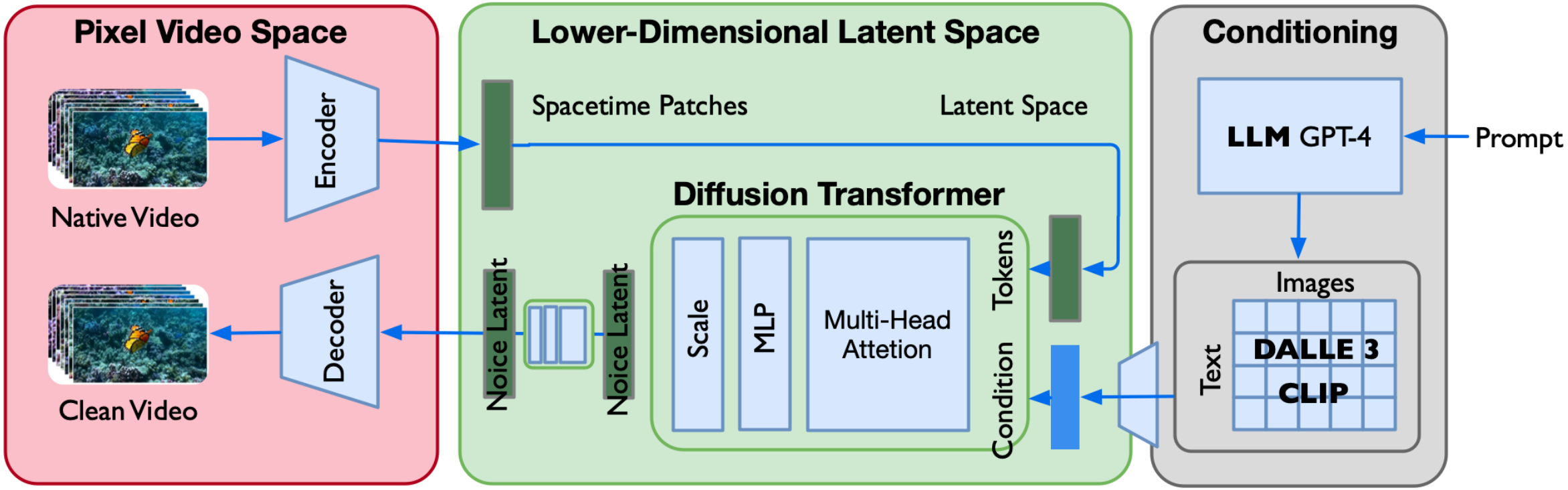
相关技术架构



SORA 模型结构

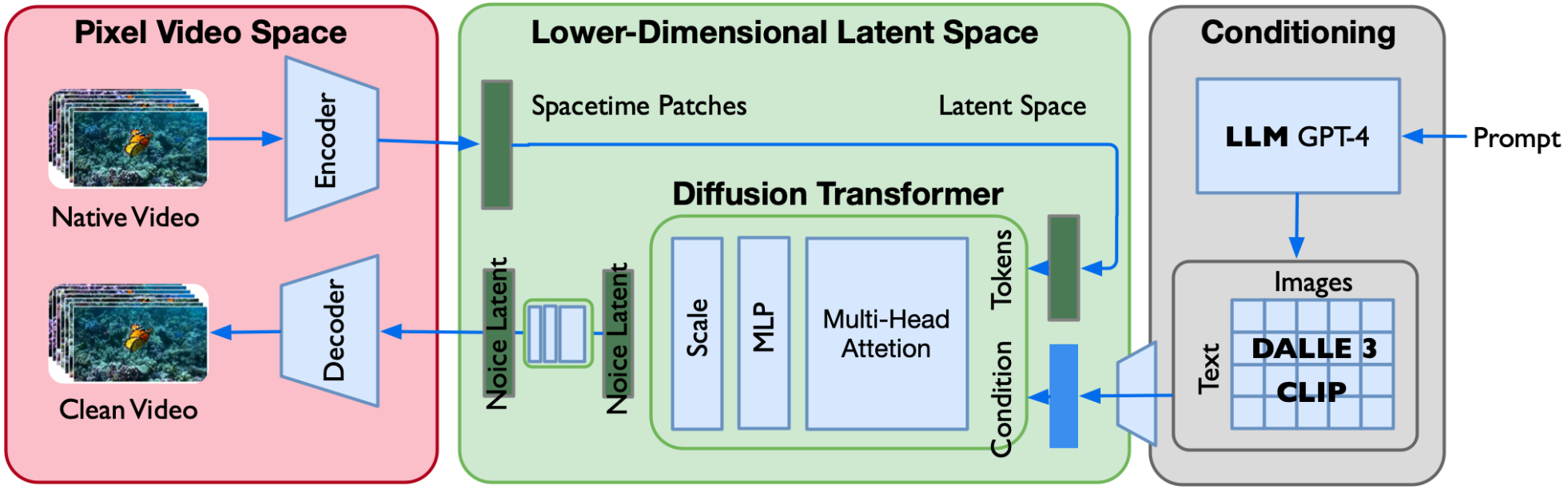
- SORA 模型结构可以表示：

$$\text{SORA} = [\text{VAE encoder} + \text{DiT (DDPM)} + \text{VAE decoder} + \text{CLIP}]$$



SORA 模型训练流程

- Step1：使用 DALLE 3（ CLIP ） 把文本和图像对 <text，image> 联系起来；
- Step2：视频数据切分为 Patches 通过 VAE 编码器压缩成低维空间表示；
- Step3：基于 Diffusion Transformer 从图像语义生成，完成从文本语义到图像语义进行映射；
- Step4：DiT 生成的低维空间表示，通过 VAE 解码器恢复成像素级的视频数据；



模型训练 Denoising Diffusion Probabilistic Model，DDPM

- 基于扩散模型 Diffusion Model

- 基于 Diffusion Transformer

1 模型训练：扩散模型 DDPM

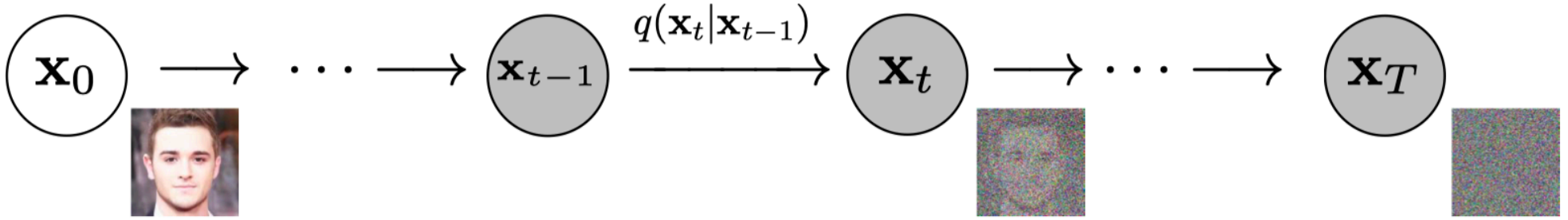


Figure 2: The directed graphical model considered in this work.

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

1 模型训练：扩散模型 DDPM

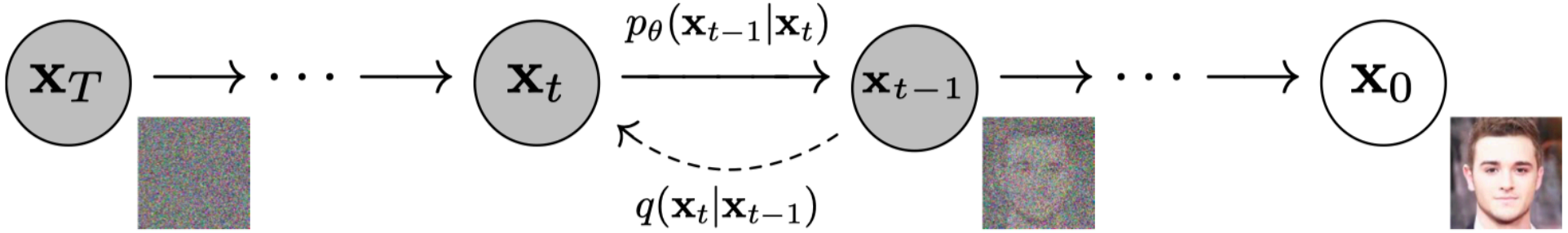


Figure 2: The directed graphical model considered in this work.

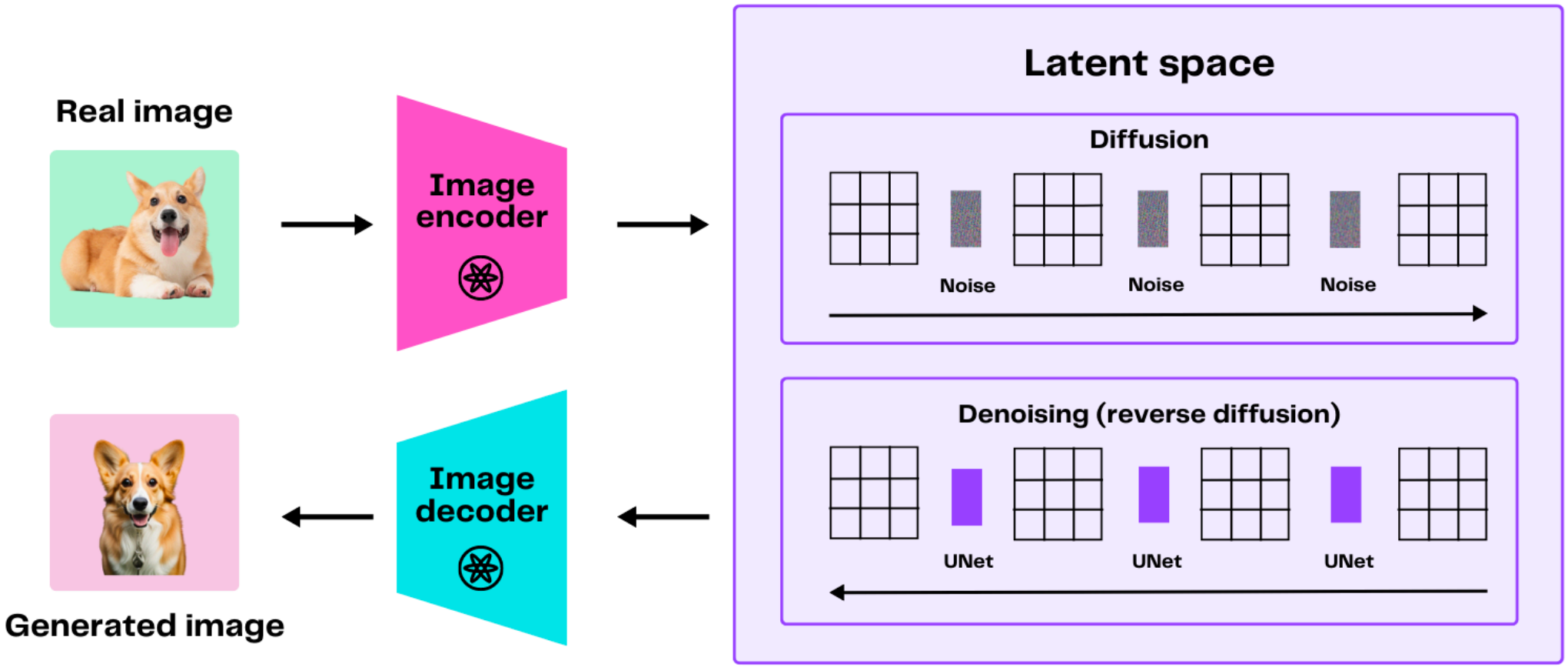
Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

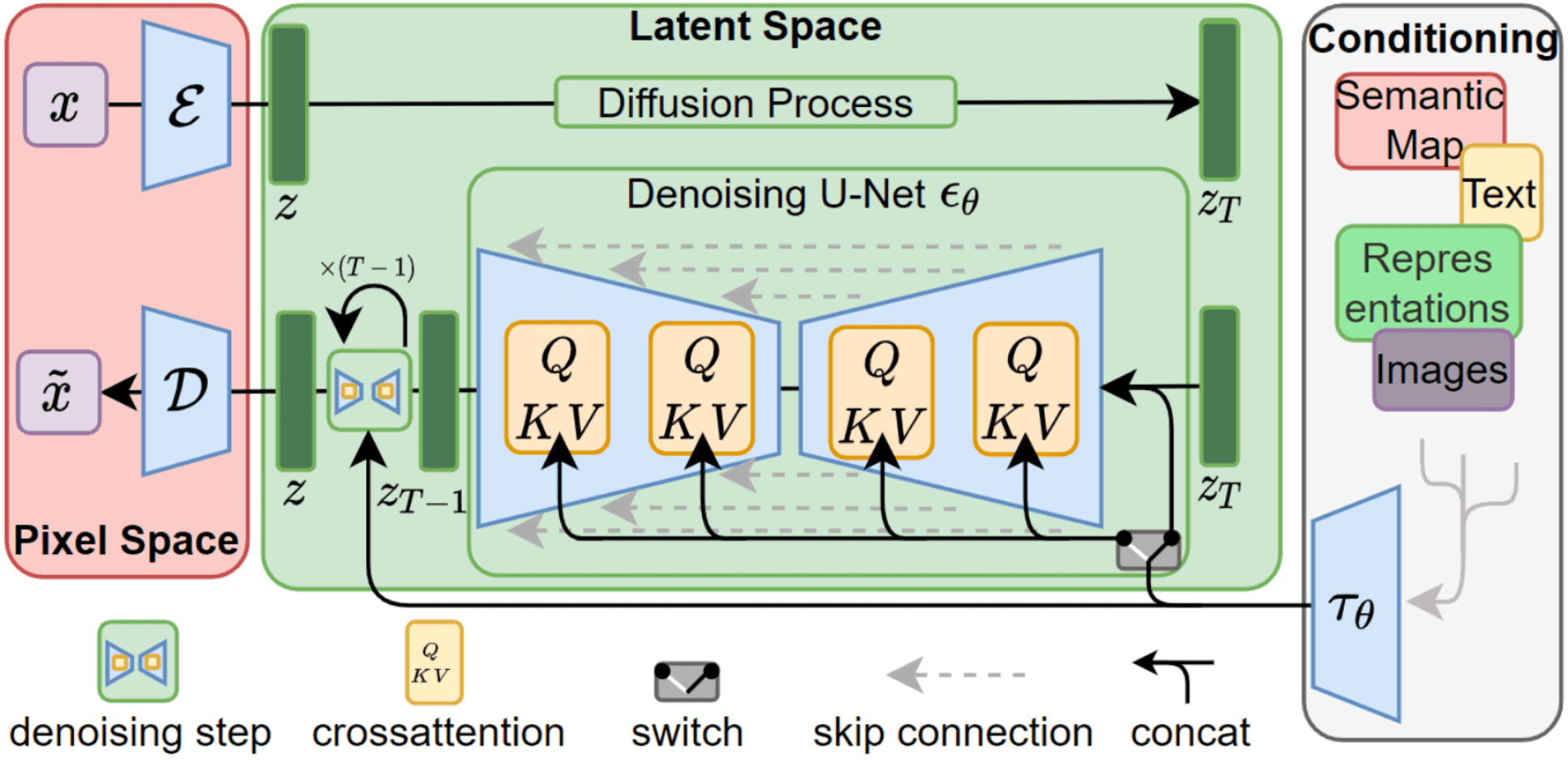
Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

1 模型训练：扩散模型 DDPM

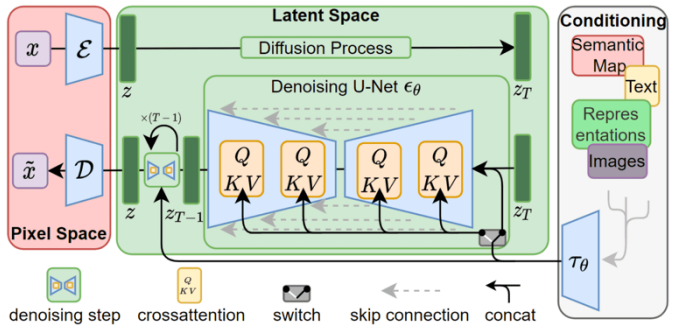


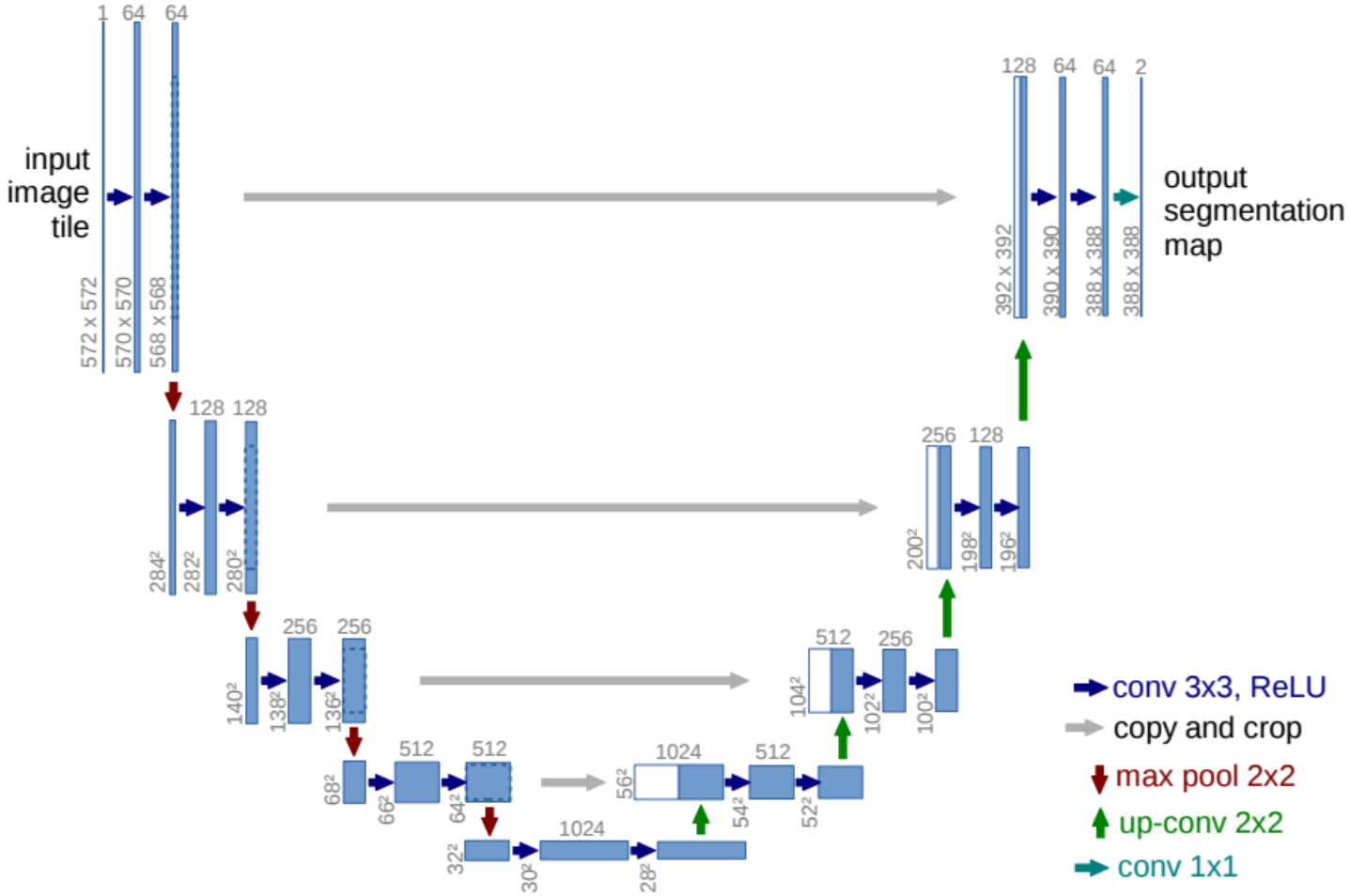
1 模型训练：基于扩散模型 SD/SDXL



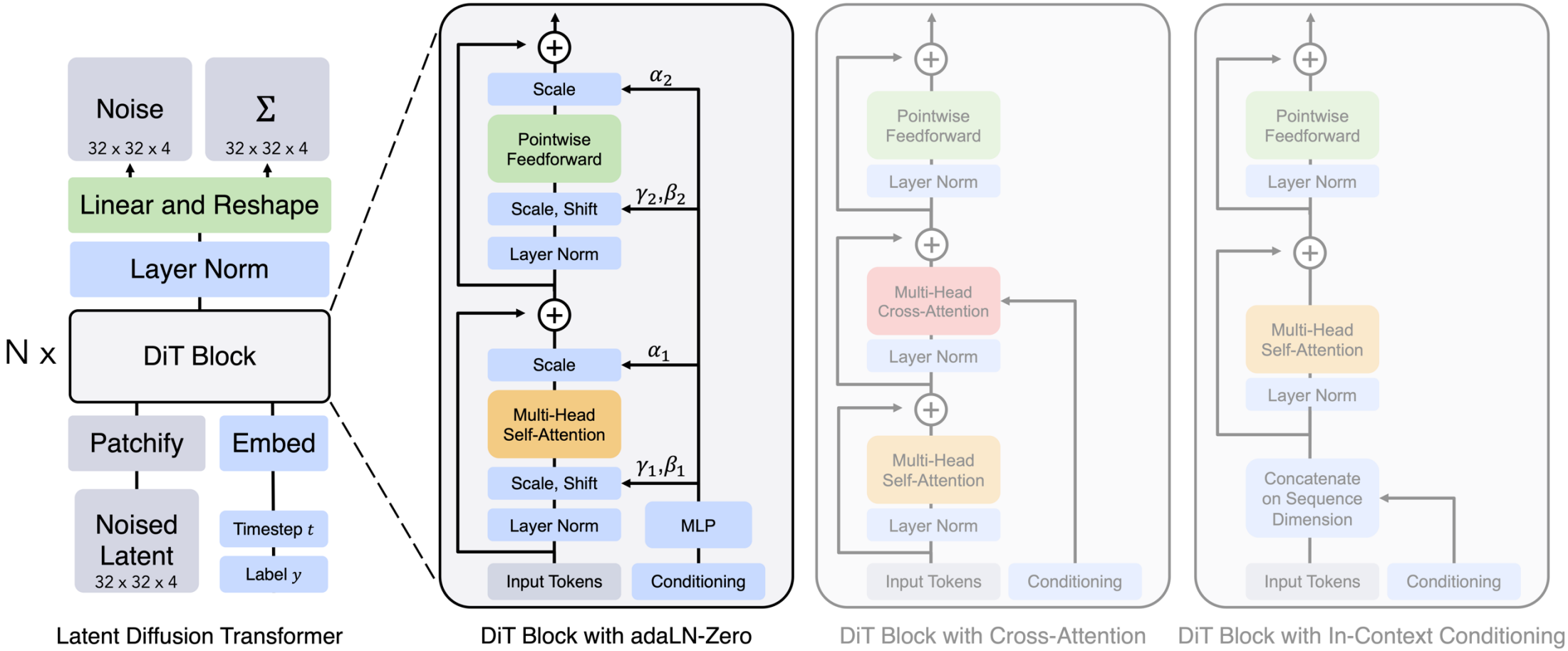
1 模型训练：基于扩散模型的主干 U-Net

1. U-Net 网络模型结构把模型规模限定；
2. SD/SDXL 作为经典网络只公布了推理和微调；
3. 国内主要基于 SD/SDXL 进行二次创作；



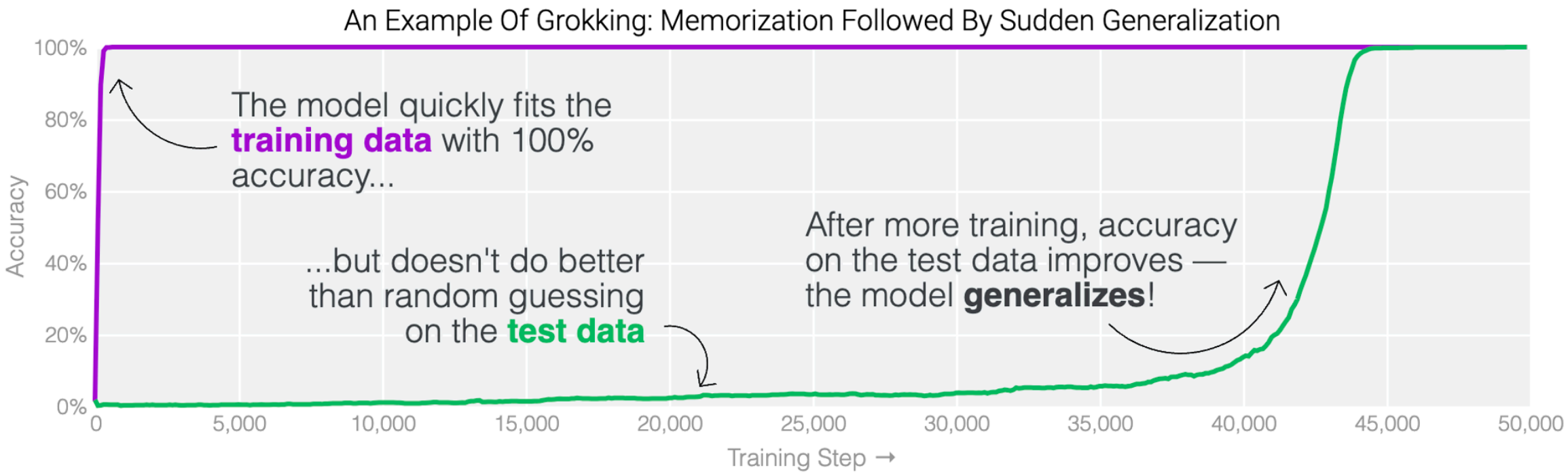


1 模型训练：基于 Diffusion Transformer

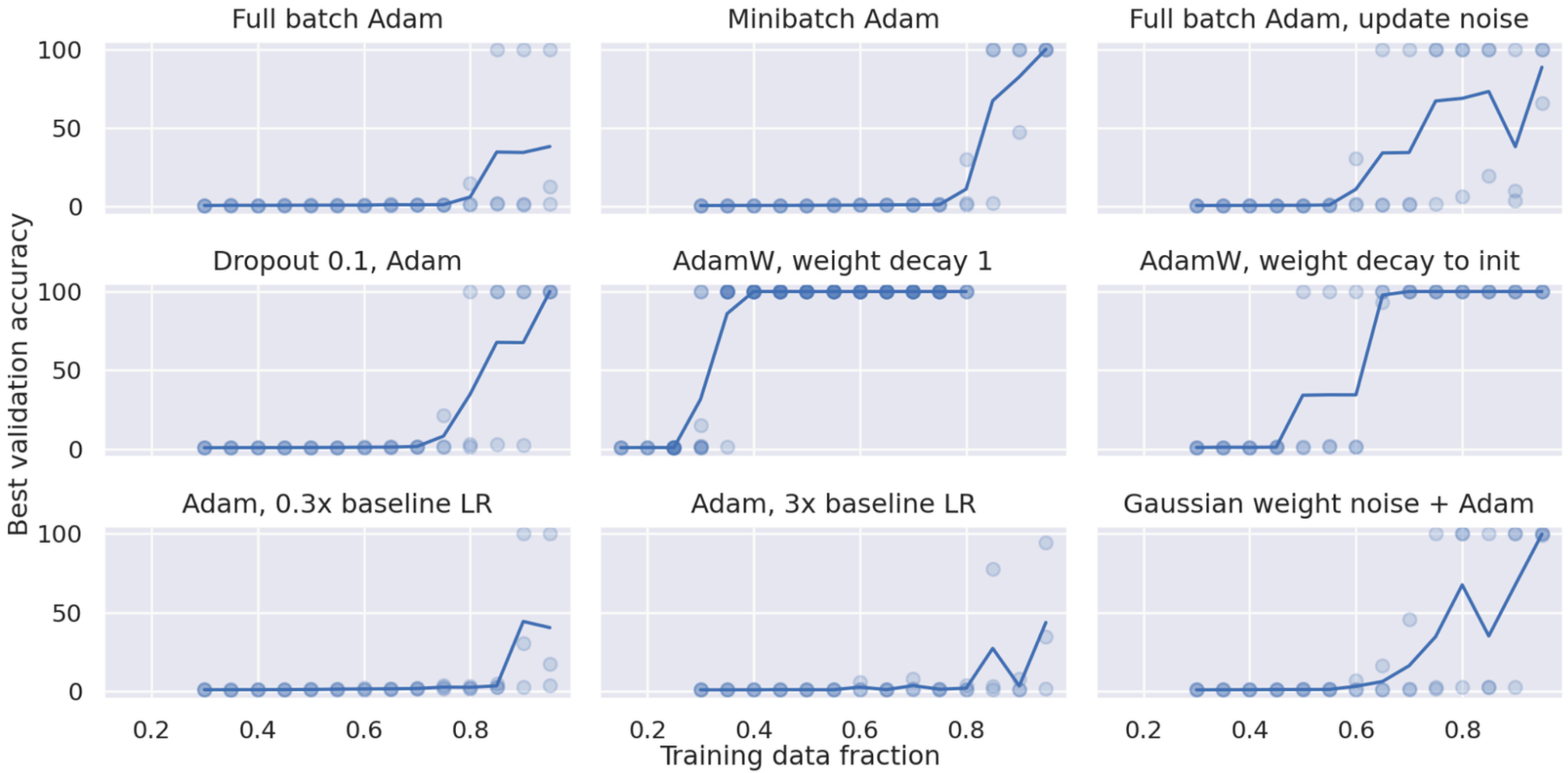


1 模型训练：大模型 Scaling Law

- GPT1->GPT2->GPT3，模型参数量从1亿到1750亿，效果产生质的变化；
- 扩大视频生成模型参数规模，迈向创建能够模拟物理世界的通用工具有前途的一步。

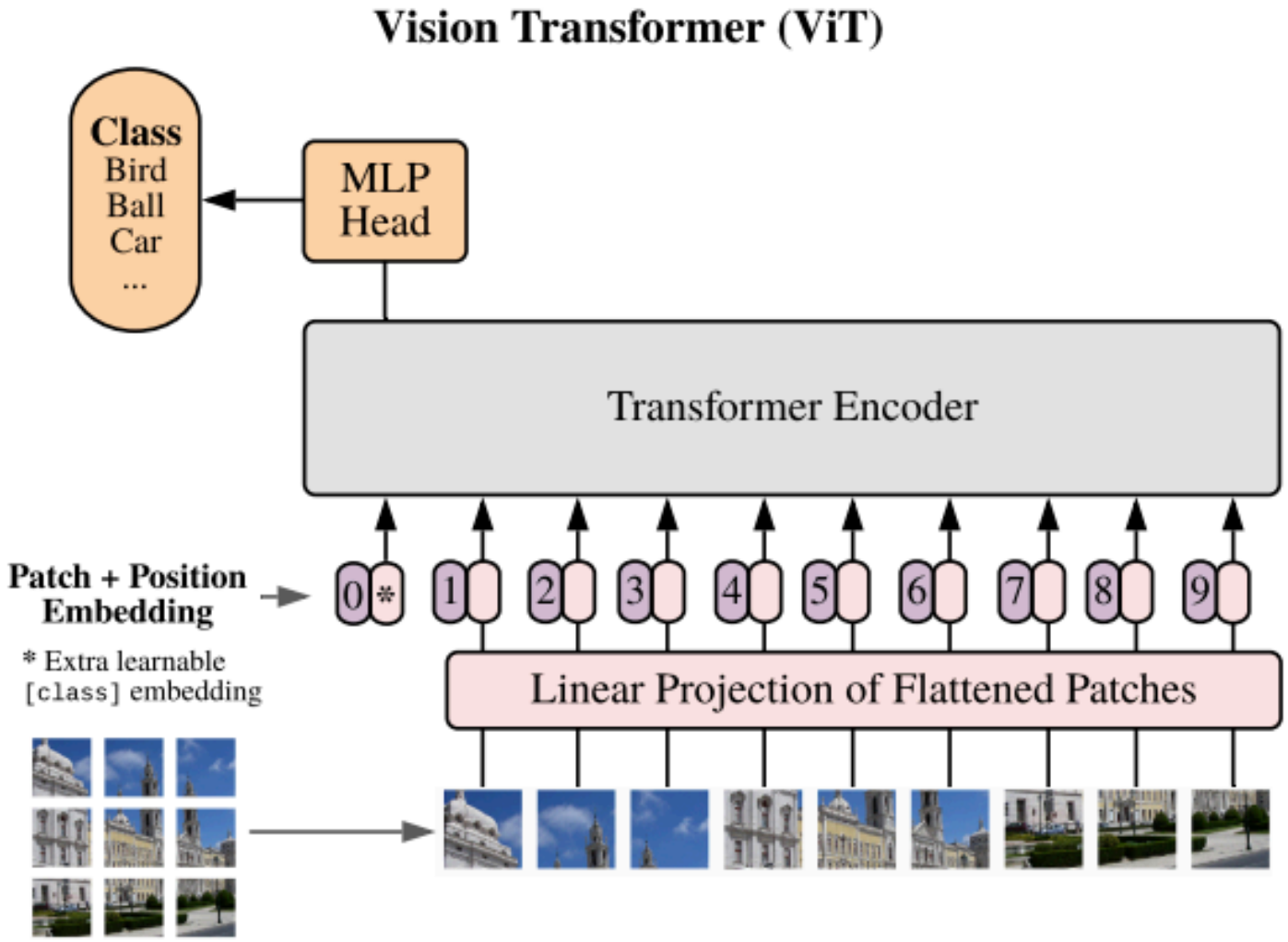


1 模型训练：大模型 Scaling Law



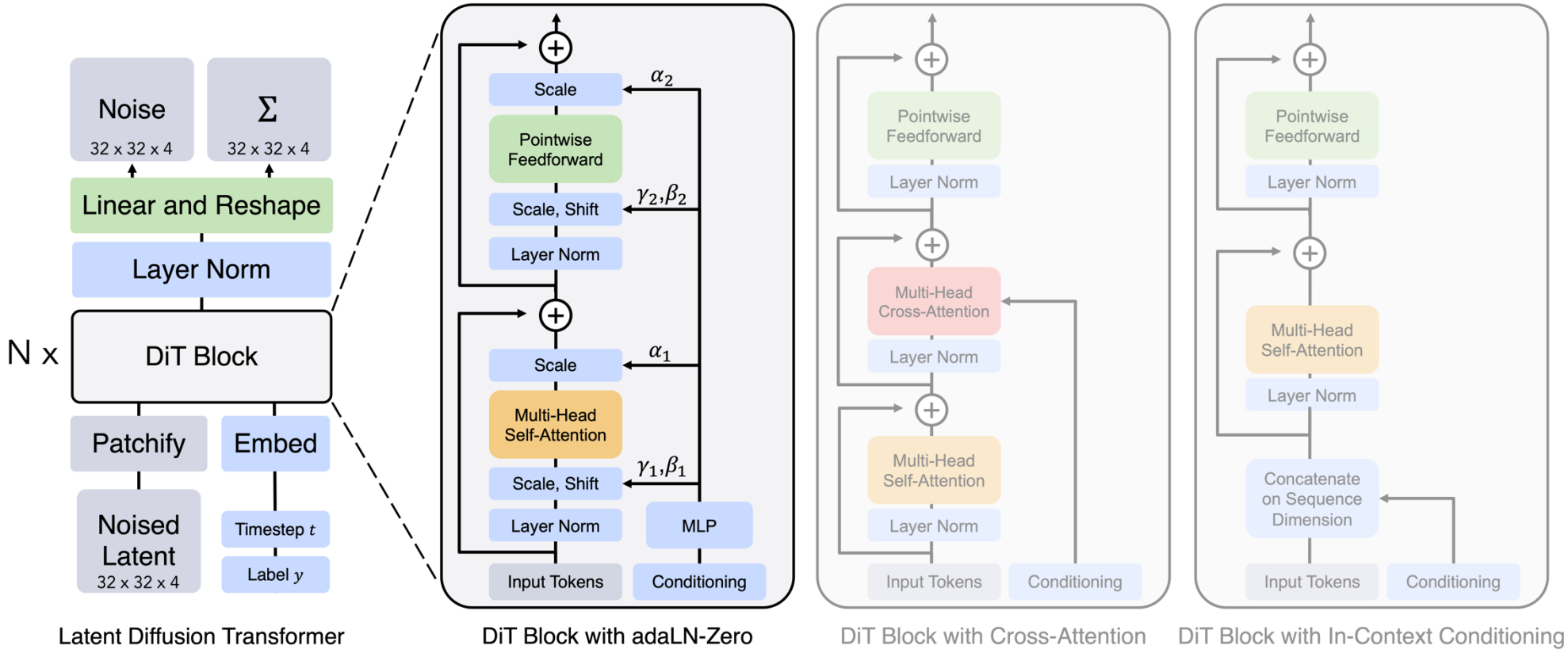
2 网络结构：Vision Transformer，ViT

- ViT 尝试将标准 Transforme r 结构直接应用于图像；
- 图像被划分为多个 patch 后，将二维 patch 转换为一维向量作为 Transformer 的输入；



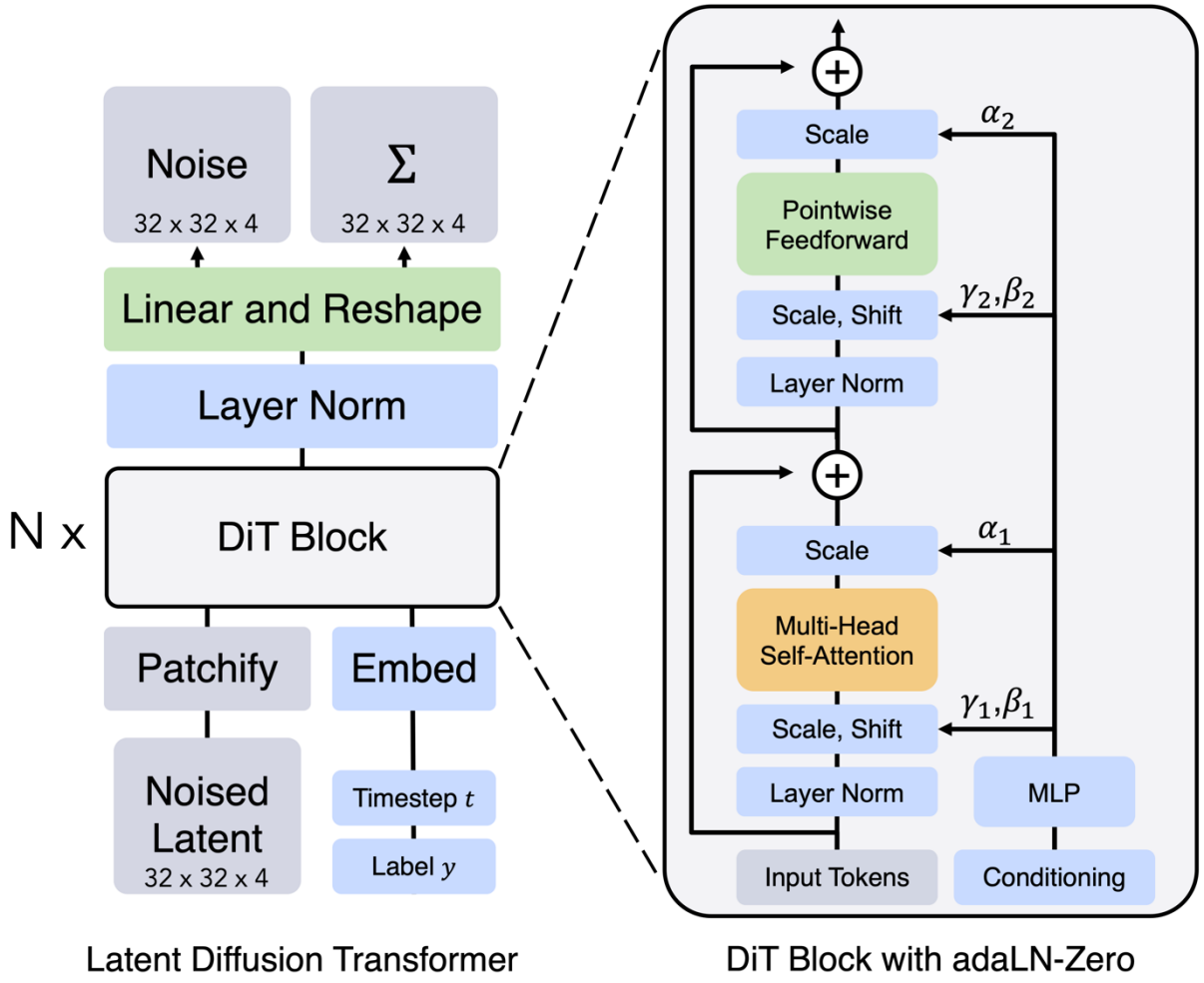
2 网络结构：Diffusion Transformer，DiT

- DiT 利用 transformer 结构探索新的扩散模型，成功用 transformer 替换 U-Net 主干；



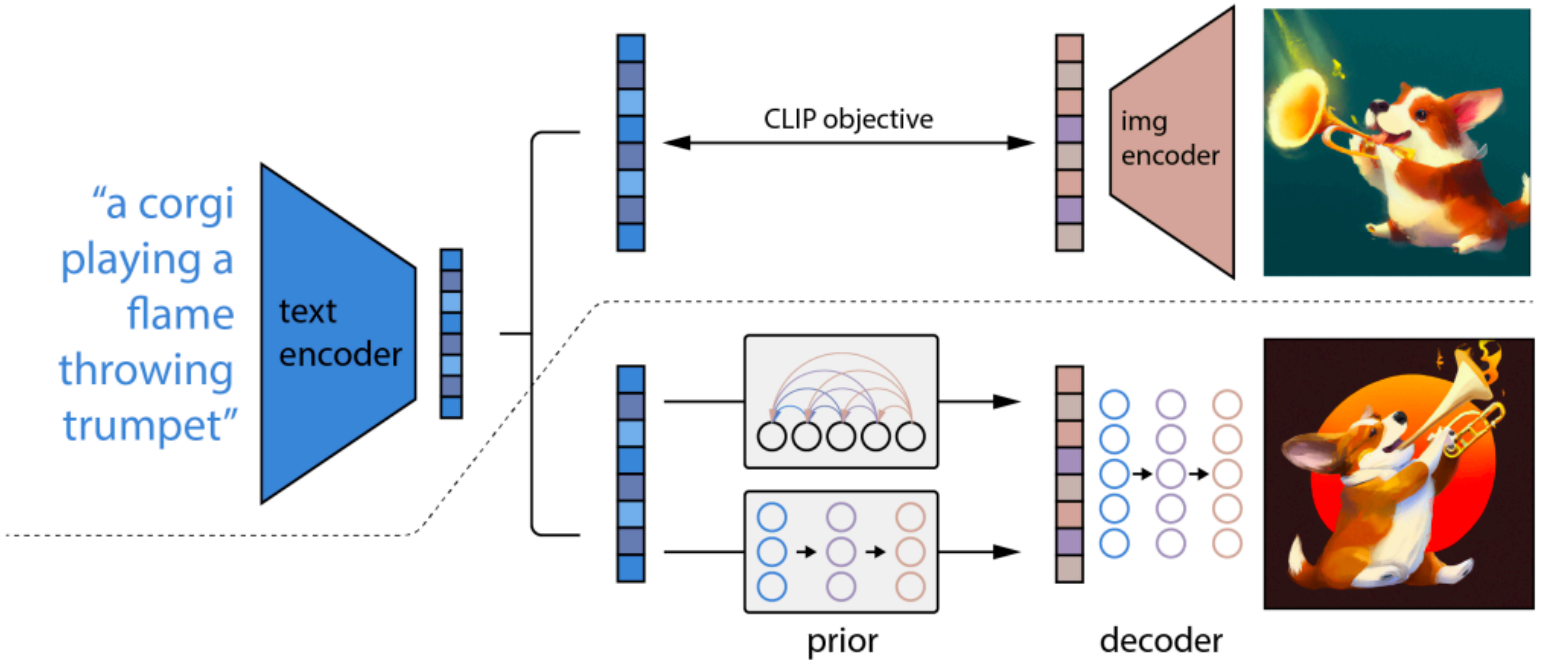
2 网络结构：Diffusion Transformer，DiT

- DiT 首先将将每个 patch 空间表示 Latent 输入到第一层网络，以此将空间输入转换为 tokens 序列。
- 将标准基于 ViT 的 Patch 和 Position Embedding 应用于所有输入 token，最后将输入 token 由 Transformer 处理。
- DiT 还会处理额外信息，e.g. 时间步长、类别标签、文本语义等。



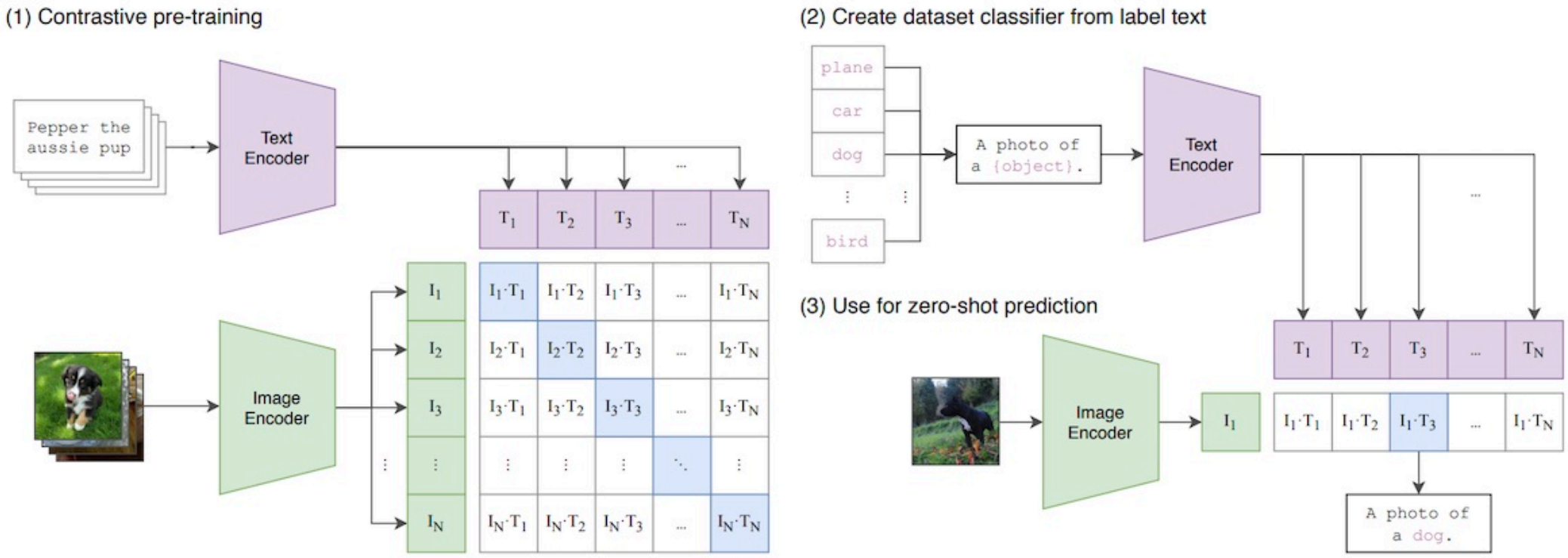
2 网络结构： DALLE 2

1. 将文本提示输入文本编码器，该训练过的编码器便将文本提示映射到表示空间；
2. 先验模型将文本编码映射到图像编码，图像编码捕获文本编码中的语义信息；
3. 图像解码模型随机生成一幅从视觉上表现该语义信息的图像；



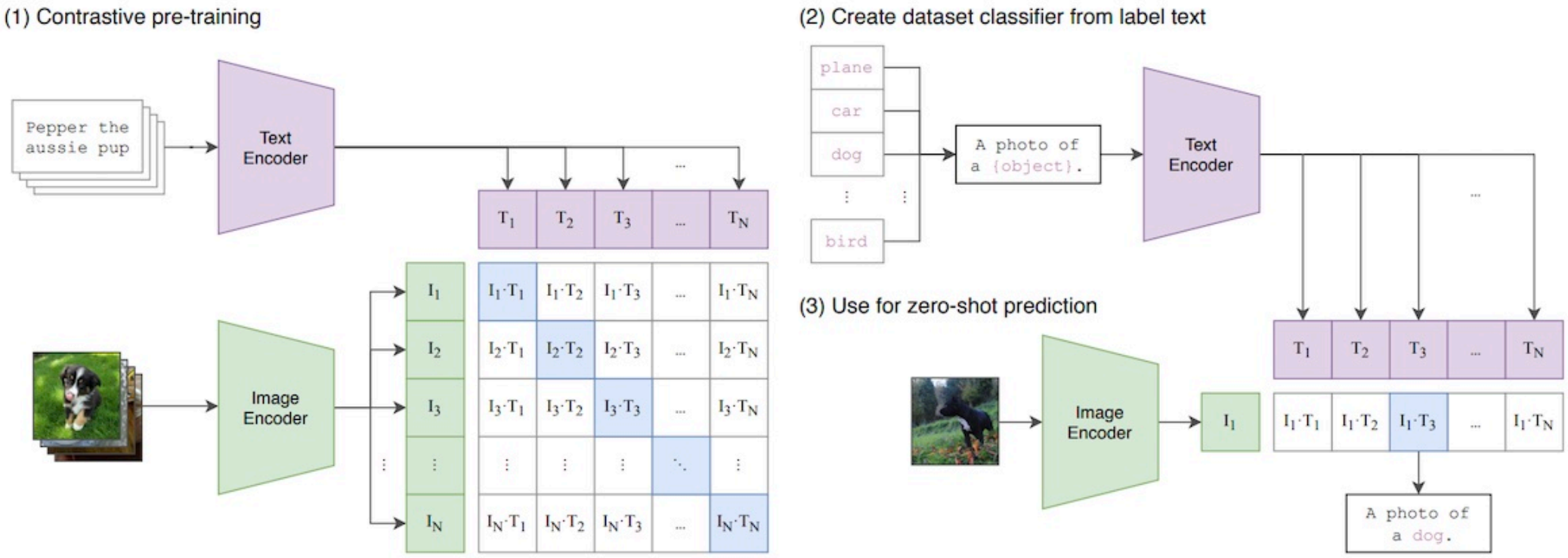
2 网络结构： CLIP

- DALL-E 2 中文本语义和与其相对的视觉图片之间的联系，由 OpenAI 模型 CLIP（**C**ontrastive **L**anguage-**I**mage **P**re-training）学习得到。



2 网络结构： CLIP

- CLIP 接受~亿对 <图片-文字> 数据训练，学习到给定文本与图像的关系；
- CLIP 并不是试图预测给定图像对应文字说明，而是学习给定文本与图像之间的关联；

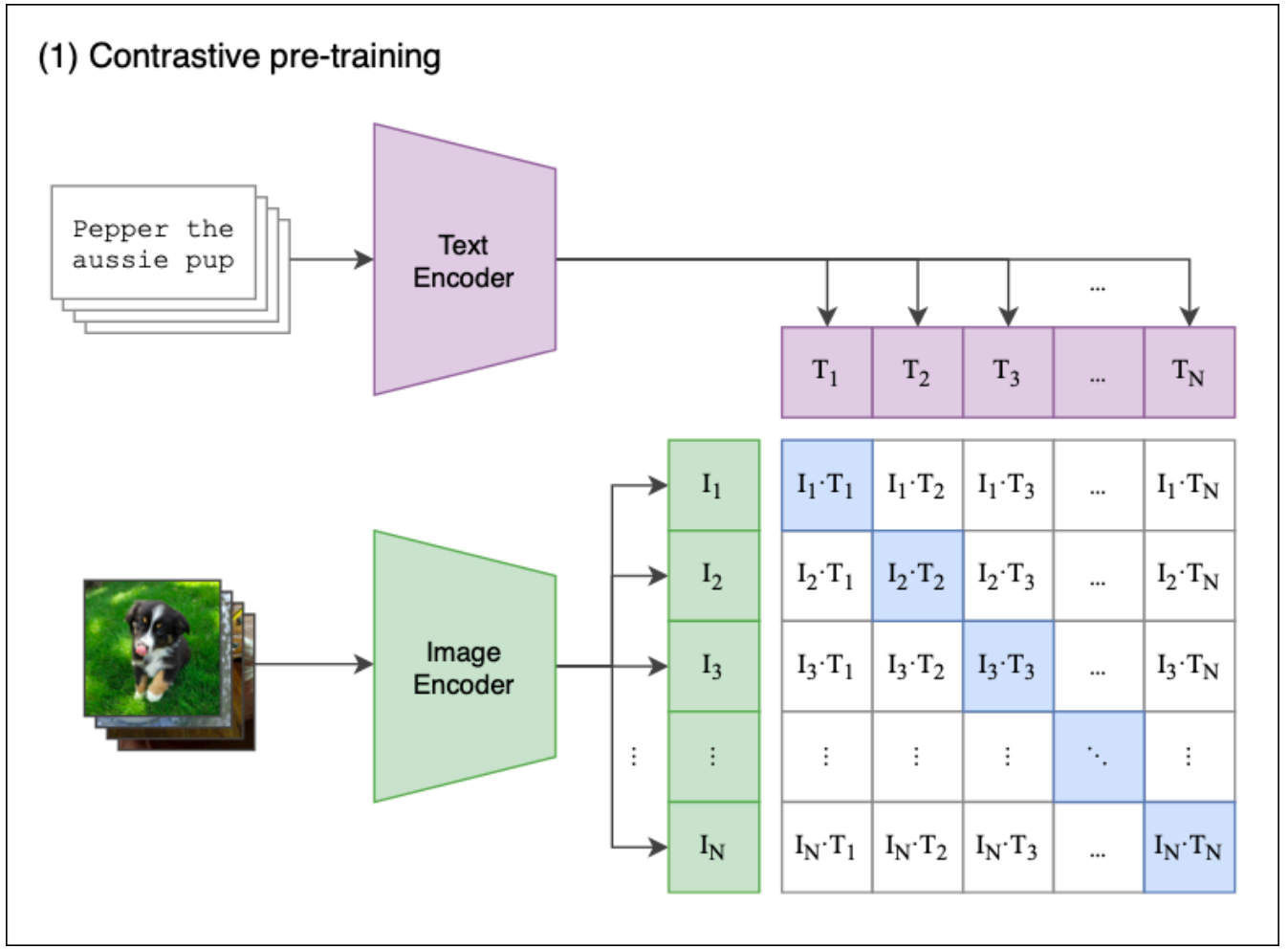


2 网络结构： CLIP

1. 图像及文本通过各自编码器，映射到 m 维空间；

2. 计算每个 <图像，文本>对的 cos 值相似度；

3. 训练使正确编码 <图像，文本> 间cos值相似度最大化；



技术总结

1. **Scaling Law：模型规模的增大**对视频生成质量的提升具有明确意义，从而很好地解决视频一致性、连续性等问题；
2. **Data Engine：数据工程很重要**，如何设计视频的输入（e.g. 是否截断、长宽比、像素优化等）、patches 的输入方式、文本描述和文本图像对质量；
3. **AI Infra：AI 系统（AI 框架、AI 编译器、AI 芯片、大模型）工程化**能力是很大的技术壁垒，决定了 Scaling 的规模。
4. **LLM：**LLM 大语言模型仍然是核心，多模态（文生图、图生文）都需要文本语义去牵引和约束生成的内容，CLIP/BLIP/GLIP 等关联模型会持续提升能力；

思考 & 总结

思考点

1. **算力**：对算力需求增长如何？如 LLM 在服务器形态爆发？推理生产应用端爆发增长？
2. **厂商**：对国产芯片和厂商带来哪些思考？需要快速复现跟进？
3. **应用**：哪些科技公司受益？哪些科技公司会持续跟进？市场会有哪些新的变化？
4. **个人**：AI 行业变化如此之快，个人应该如何转型和抓住时代机遇？

思考点 1：算力需求增长

- **对算力需求增长如何？如 LLM 在服务器形态爆发？推理生产应用端爆发增长？**

1. SORA 模型参数量预计 <10B，模型参数量不会像 LLM 需要千卡/万卡大规模 AI 集群训练（~百卡）；
2. DALL·E 3 视频文本标注数据有限（<30B），训练数据不像 LLM 可以无监督学习；
3. OpenAI 尚未公布 SORA 商业化时间，视频生成距离成熟还有时间距离（< 半年）；
4. 技术上输入内容控制一致性等问题仍需解决，推理算力全面爆发仍然有时间差（> 半年）；
5. 目前推理算力比 SD、SDXL 要大2/3个量级，需要结合 AI 训练集群或者 AI 推理集群；
6. LLM 大语言模型仍然是24年消耗算力大头，多模态很多工作建立在语言之上；

思考点 2：对厂商的挑战

- **对国产芯片和厂商带来哪些思考？需要快速复现跟进？**
  1. 视频生成不像 LLM 已经统一范式，视频大模型训练场景变化快速，不再是搞定 LLAMA 系列；
  2. 小模型时代动态 Shape、PyTorch 新特性兼容、分布式并行能力、CV 算子补齐等问题开始凸显；
  3. 对 AI 编译器挑战仍然非常大， 如何提供如 CUDA 般灵活编程体系并且拓展应用生态；
  4. 快速兼容 PT 三方生态，e.g. 只支持头部三方库如 Huggingface、Megatron-LM 等远不能满足；
  5. CPU 的图像/视频处理会开始往 GPU 落地，如 NV DALI 等图像编码加速库；

思考点 3：对市场策略的思考

- **哪些科技公司受益？哪些科技公司会持续跟进？市场会有哪些新的变化？**

1. 从底层改变内容生产方式，60秒高保真视频生成，多模态发展迎来新阶段，商用空间有望大幅打开；
2. 大力出奇迹：多模态大模型训练及应用普及对算力消耗将继续增长，看好 NVIDIA、AMD 芯片厂商；
3. 用户基础平台：基于已有用户做转化，可提升付费率及ARPPU，看好 Adobe、美图、焦点等；
4. 创业公司差距拉大：大模型技术迭代快，基于 LLAMA 微调百模大战可能迎来新一轮洗牌；

思考点 4：个人的发展

- AI 行业变化如此之快，个人应该如何转型和抓住时代机遇？

引用

1. Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhudinov. "Unsupervised learning of video representations using lstms." International conference on machine learning. PMLR, 2015.
2. Chiappa, Silvia, et al. "Recurrent environment simulators." arXiv preprint arXiv:1704.02254 (2017).
3. Ha, David, and Jürgen Schmidhuber. "World models." arXiv preprint arXiv:1803.10122 (2018).
4. Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." Advances in neural information processing systems 29 (2016).
5. Tulyakov, Sergey, et al. "Mocogan: Decomposing motion and content for video generation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
6. Clark, Aidan, Jeff Donahue, and Karen Simonyan. "Adversarial video generation on complex datasets." arXiv preprint arXiv:1907.06571 (2019).
7. Brooks, Tim, et al. "Generating long videos of dynamic scenes." Advances in Neural Information Processing Systems 35 (2022): 31769-31781.
8. Yan, Wilson, et al. "Videogpt: Video generation using vq-vae and transformers." arXiv preprint arXiv:2104.10157 (2021).
9. Wu, Chenfei, et al. "Nüwa: Visual synthesis pre-training for neural visual world creation." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
10. Ho, Jonathan, et al. "Imagen video: High definition video generation with diffusion models." *arXiv preprint arXiv:2210.02303* (2022).
11. Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
12. Gupta, Agrim, et al. "Photorealistic video generation with diffusion models." arXiv preprint arXiv:2312.06662 (2023).
13. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
14. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
15. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
16. Arnab, Anurag, et al. "Vivit: A video vision transformer." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
17. He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
18. Dehghani, Mostafa, et al. "Patch n'Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution." *arXiv preprint arXiv:2307.06304* (2023).
19. Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
20. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).
21. Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *International conference on machine learning*. PMLR, 2015.
22. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.
23. Nichol, Alexander Quinn, and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models." *International Conference on Machine Learning*. PMLR, 2021.
24. Dhariwal, Prafulla, and Alexander Quinn Nichol. "Diffusion Models Beat GANs on Image Synthesis." *Advances in Neural Information Processing Systems*. 2021.
25. Karras, Tero, et al. "Elucidating the design space of diffusion-based generative models." *Advances in Neural Information Processing Systems* 35 (2022): 26565-26577.
26. Peebles, William, and Saining Xie. "Scalable diffusion models with transformers." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
27. Chen, Mark, et al. "Generative pretraining from pixels." *International conference on machine learning*. PMLR, 2020.
28. Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International Conference on Machine Learning*. PMLR, 2021.
29. Yu, Jiahui, et al. "Scaling autoregressive models for content-rich text-to-image generation." *arXiv preprint arXiv:2206.10789* 2.3 (2022): 5.
30. Betker, James, et al. "Improving image generation with better captions." *Computer Science. https://cdn.openai.com/papers/dall-e-3. pdf* 2.3 (2023): 8
31. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." *arXiv preprint arXiv:2204.06125* 1.2 (2022): 3.
32. Meng, Chenlin, et al. "Sdedit: Guided image synthesis and editing with stochastic differential equations." *arXiv preprint arXiv:2108.01073* (2021).
33. An Initial Exploration of Theoretical Support for Language Model Data Engineering. Part 1: Pretraining