

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 11.1: Generation - Conditioning

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

Week 11 Reading Assignment (due Friday 11/6 at 8pm ET)

- Two papers (with a focus on multimodal generation)

Week 12 Reading Assignment (due Friday 11/13 at 8pm ET)

- Two papers (with a focus on multimodal generation)

Final project presentations (Tuesday 12/1 and Thursday 12/3)

- Read final report instructions first!
- Poster presentations
  - Poster will focus on a subset of the research ideas and results

Final project report (due Sunday 12/6 at 8pm ET)

- N-1 research ideas, exploring new modeling approaches
  - All ideas should be detailed in the report

Advanced Multimodal Machine Learning (11-877)

11-877 AMML

home schedule syllabus projects



Advanced Topics in MultiModal Machine Learning

11-877 • Spring 2024 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, vision, and acoustic. This research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course is designed to be a graduate-level course covering recent research papers in multimodal machine learning, including technical challenges with representation, alignment, reasoning, generation, co-learning and quantification. The main goal of the course is to increase critical thinking skills, knowledge of recent technical achievements, and understanding of future research directions.

https://multicomp.cs.cmu.edu/adv-mmml-course/spring2024/

Attendance using Poll Everywhere



1. Scan QR Code
   or
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

❌ *Do not try to sign-in*

3. Click ”Share location”



Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 11.1: Generation – Conditioning

Louis-Philippe Morency

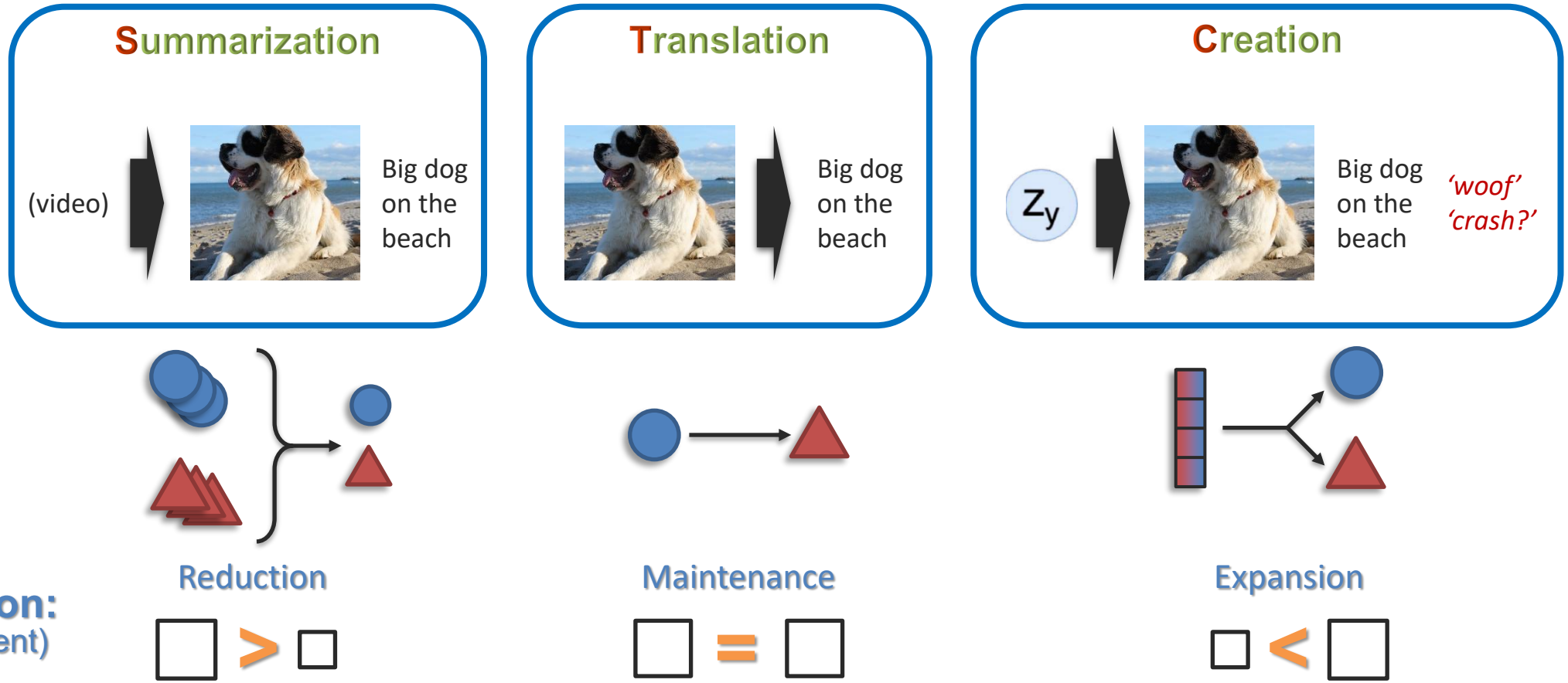
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

- Generation challenges: conditioning and coherence
  - Text conditioning of visual generation (link with Week 6)
- Diffusion: corruption, denoising network, generation process
- Adding text conditioning to a visual diffusion model
  - Latent diffusion and autoencoders
  - Diffusion Transformers: in-context, cross-attention, AdaLN
  - Multimodal diffusion transformers
  - Classifier-free guidance

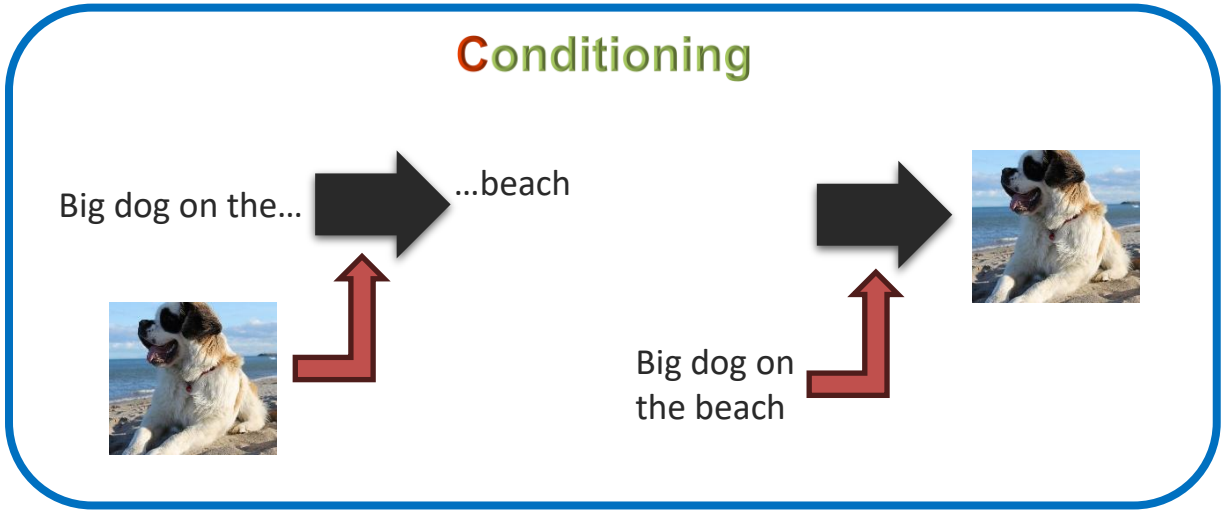
Generation v1

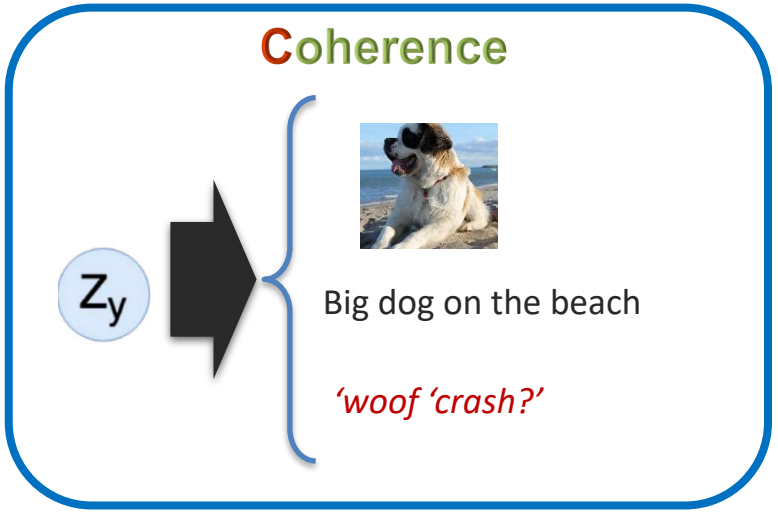
Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



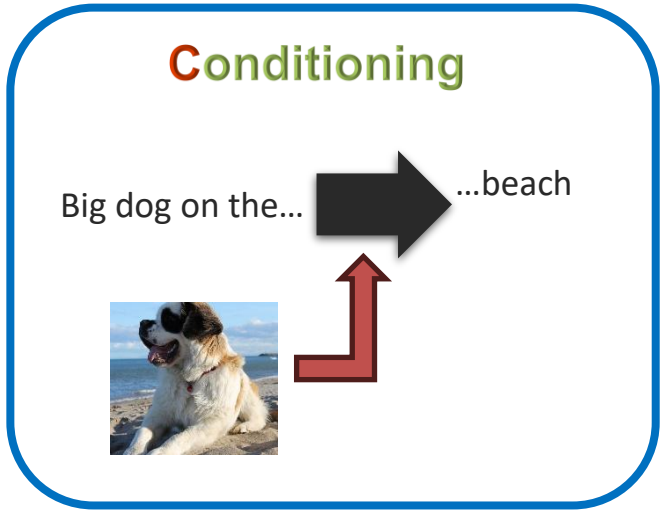
Generation v2

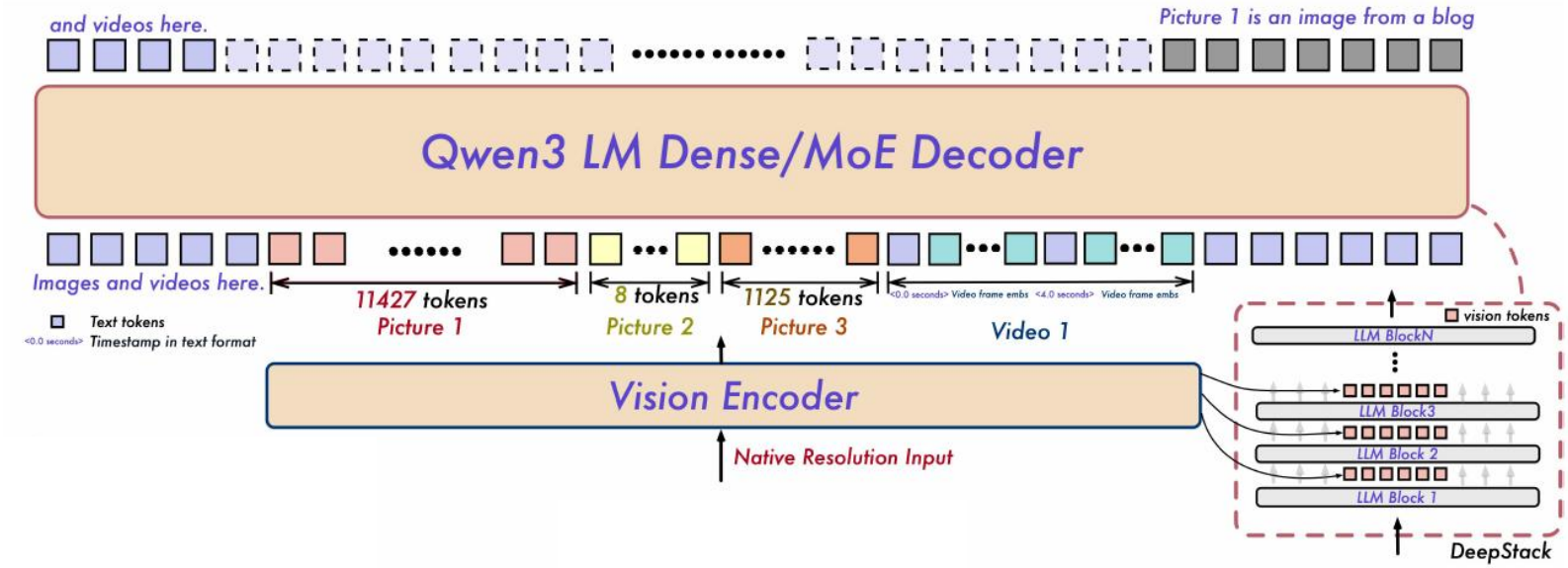
Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.





Week 6: Vision-Language Models and Multimodal LLMs





This Lecture: Text Conditioning of Visual Generation

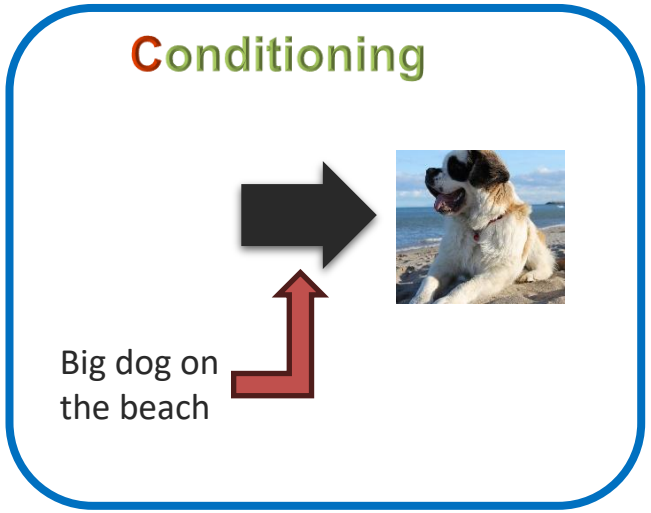




Figure 1: Generated depictions of the scene "A robot and its pet in a tree." Many elements are underspecified in the text, e.g., pet type, perspective, and visual style.

[Hutchinson, Baldridge & Prabhakaran, Underspecification in Scene Description-to-Depiction Tasks. AACL-IJCNLP 2022]

Diffusion

How to Generate Photographic Images?

x: an image



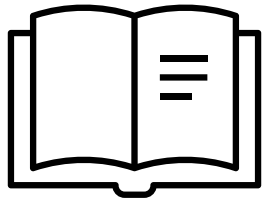
$p(x)$: probability of image being "photographic", natural

How to sample?

Idea: Learn a function to improve the quality of a "corrupted" image

But how to get paired data (corrupted and clean images?)

x: a text document



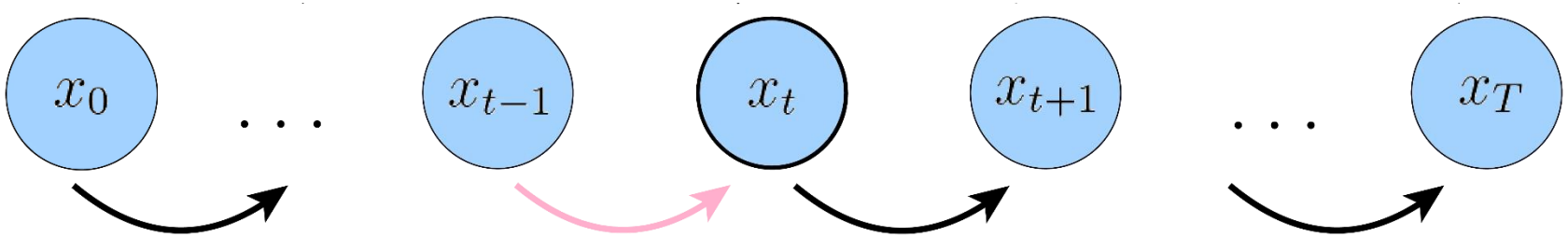
$p(x)$: probability of document being "natural" English

Autoregressive language models

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

Diffusion: Corrupting Images to Learn how to Generate Them

Corruption: a blend of the clean image and pure noise



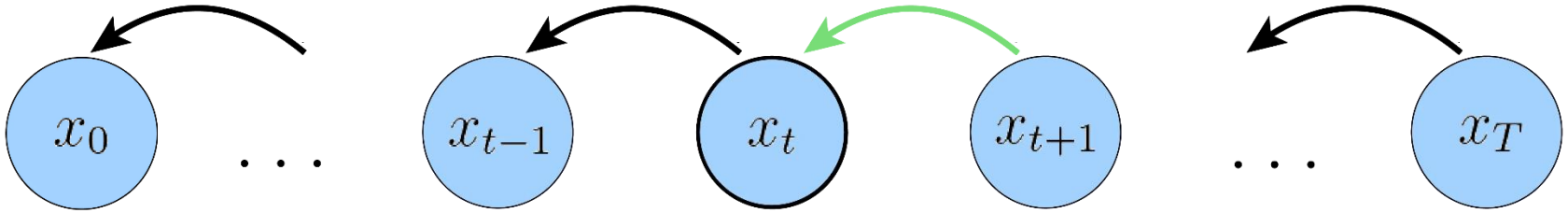


$$x_t = \underbrace{\sqrt{\bar{\alpha}_t}}_{\text{signal}}\, x_0 + \underbrace{\sqrt{1-\bar{\alpha}_t}}_{\text{noise}}\, \varepsilon \qquad \text{where } \varepsilon \sim \mathcal{N}(0, I)$$

[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

Reverse process: Image Generation

Question: Which image x_0 produced this noisy image x_t?



Key Idea: If each corruption step is small, the reverse step is also approximately Gaussian



Then, let's learn a network that can predict the noise

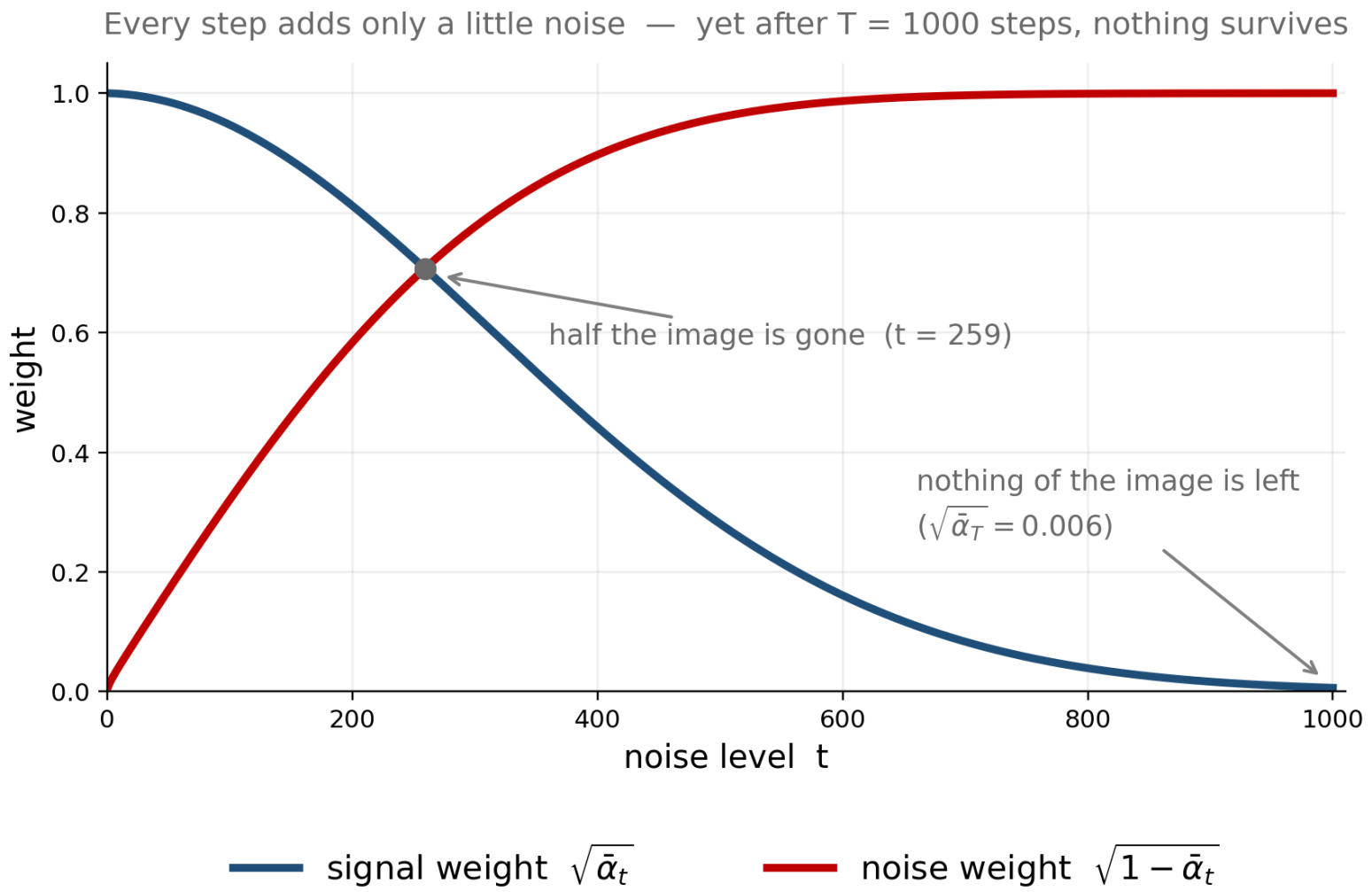
$$\hat{\varepsilon} = \varepsilon_\theta(x_t, t)$$

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\varepsilon}}{\sqrt{\bar{\alpha}_t}}$$

How to train?

[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

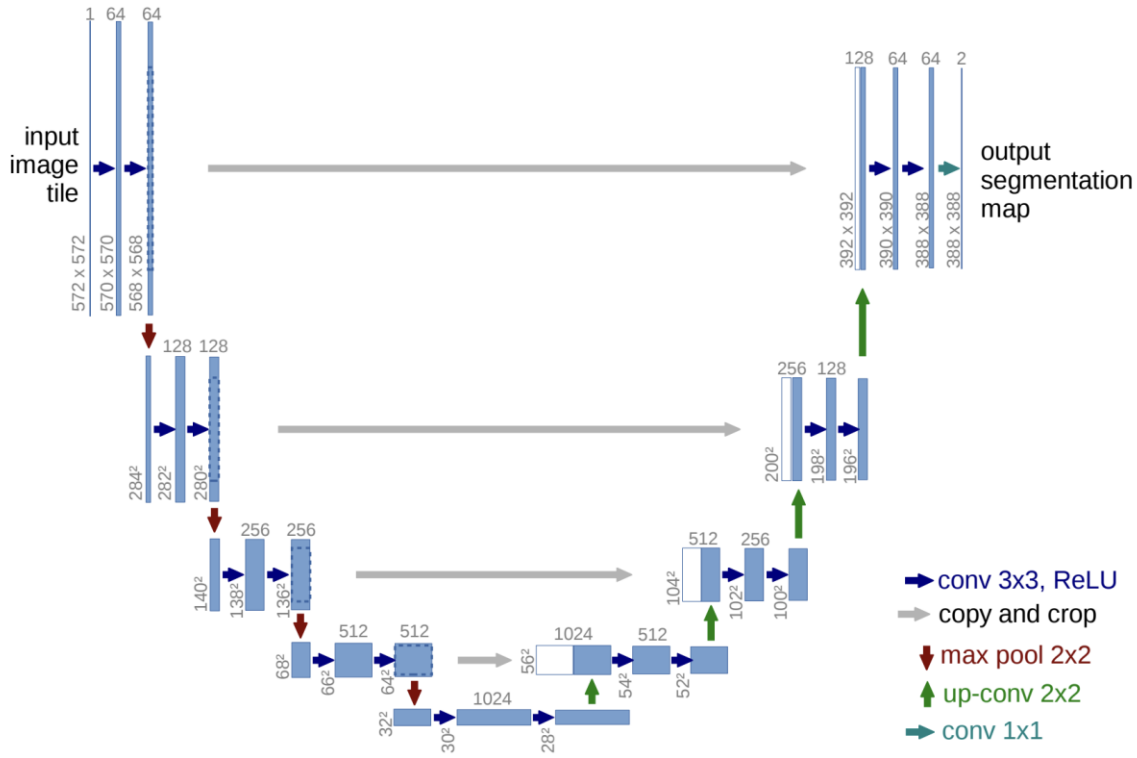
Noise Schedule



[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

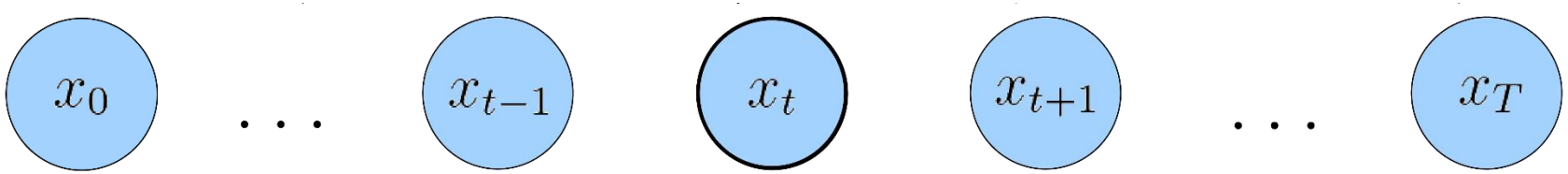
Network Implementation: U Nets

Key insight: learn transformations with convolutions, but apply them to features at the original image scale. Good for segmentation, image editing



[Ronneberger, Fischer & Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015]

Training the Denoising Network



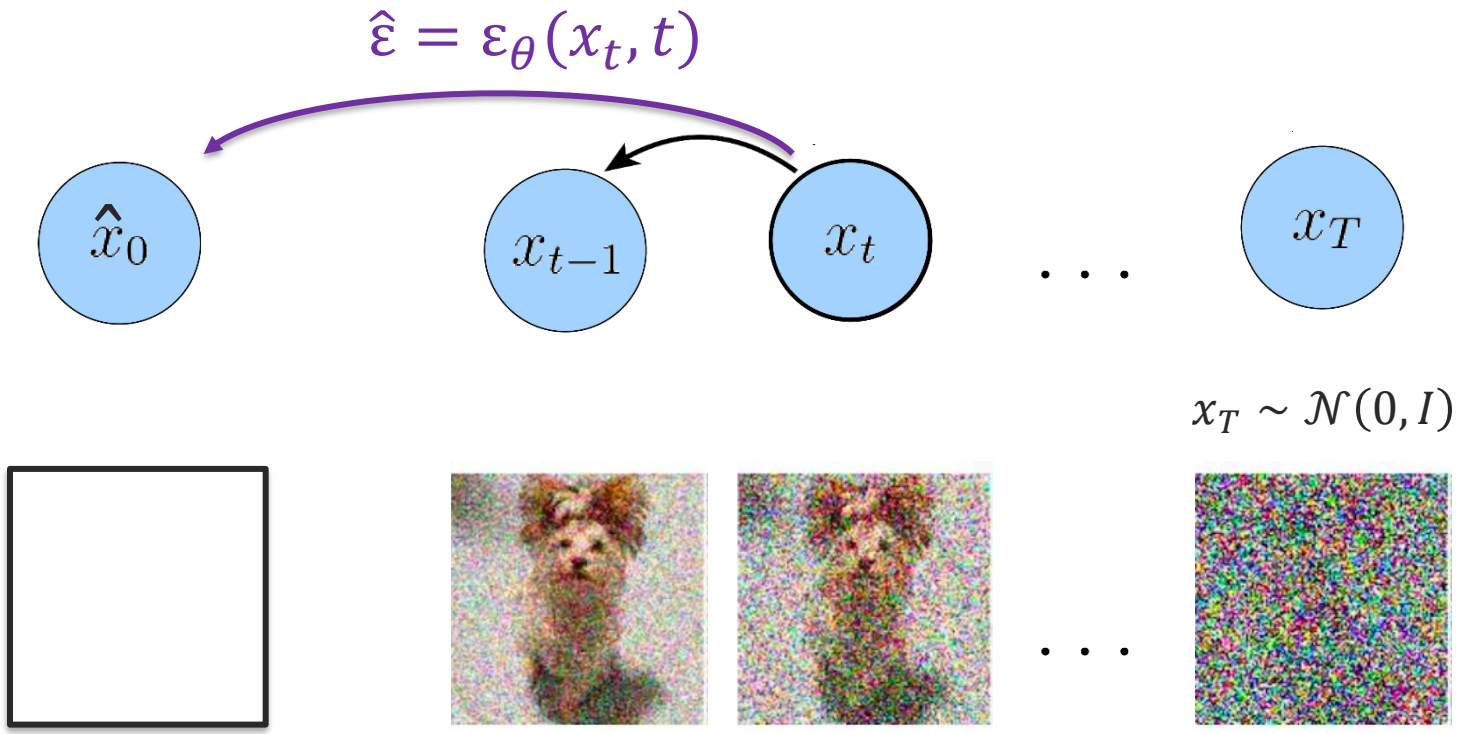


$$\mathcal{L} = \mathbb{E}_{x_0, t, \varepsilon}[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2]$$

Training: sample an image x_0 ● sample a noise level t ● sample noise ε ● corrupt image x_t ● optimize $\mathcal{L}$

[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

Image Generation Process



Remove one step's worth

Update: $$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\varepsilon}\right) + \sigma_t z$$

Re-randomize

Rescale

[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

Conditioning

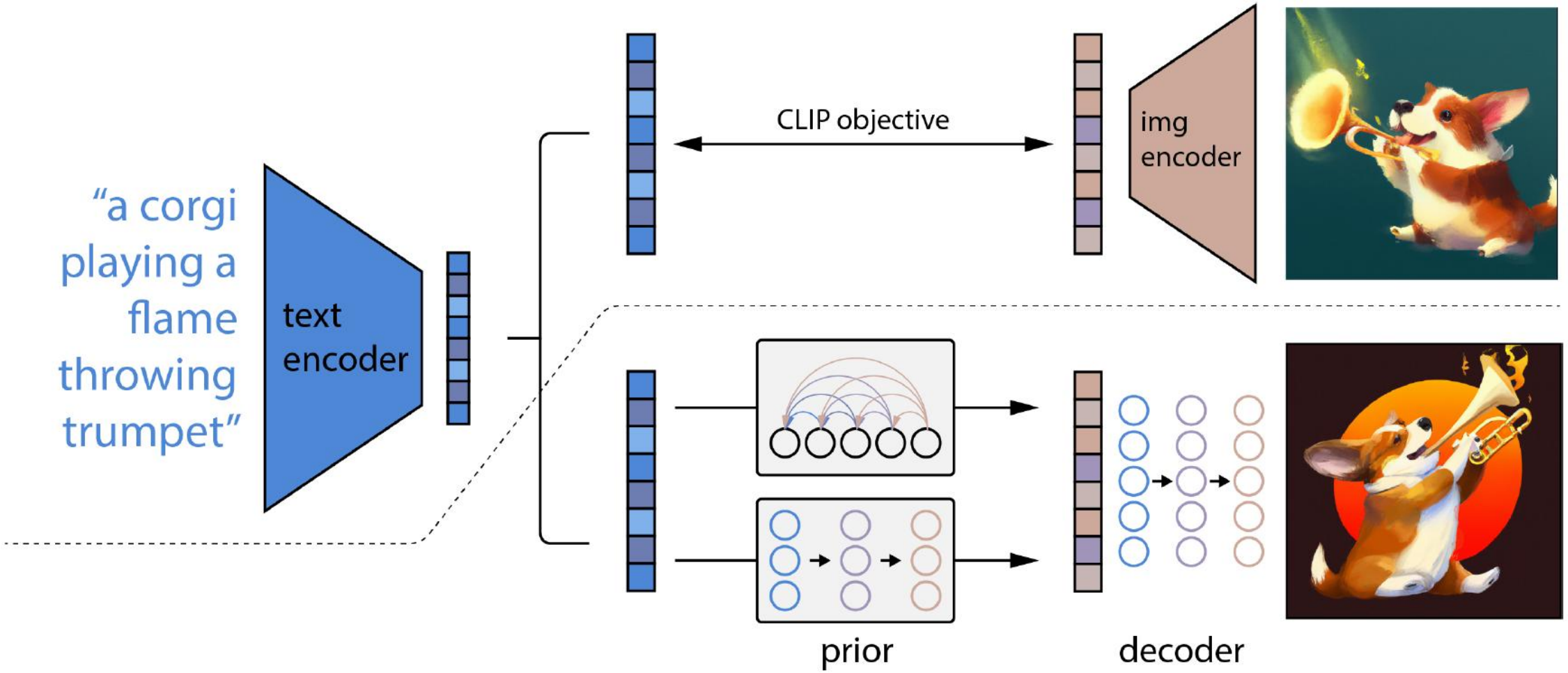
Conditional Generation

Sometimes we care about p(x|c) - **conditional generation**

- E.g., c = category (faces, outdoor images), x = image
  c = text, x = image
  c = image, x = text



Dall-E 2 (aka, unCLIP)



[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

Comparing Text Representations

DALL-E 2

Diffusion on top of frozen CLIP embeddings



✗

Imagen

Diffusion on top of frozen T5 embeddings



✓

A black apple and a green backpack.



A horse riding an astronaut.

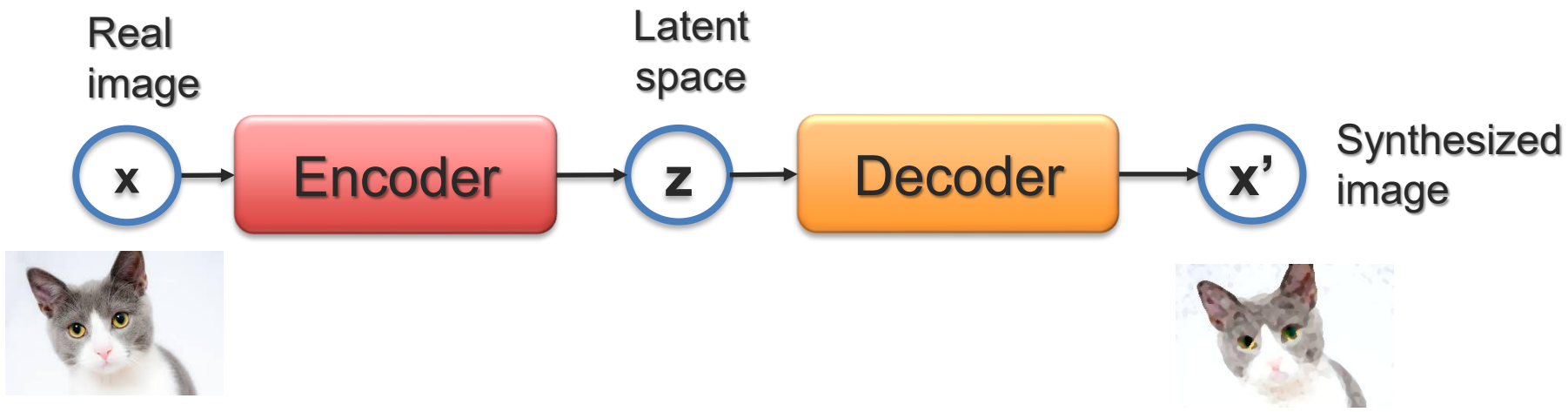
✗



✗

[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]
[Saharia et al., Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022]

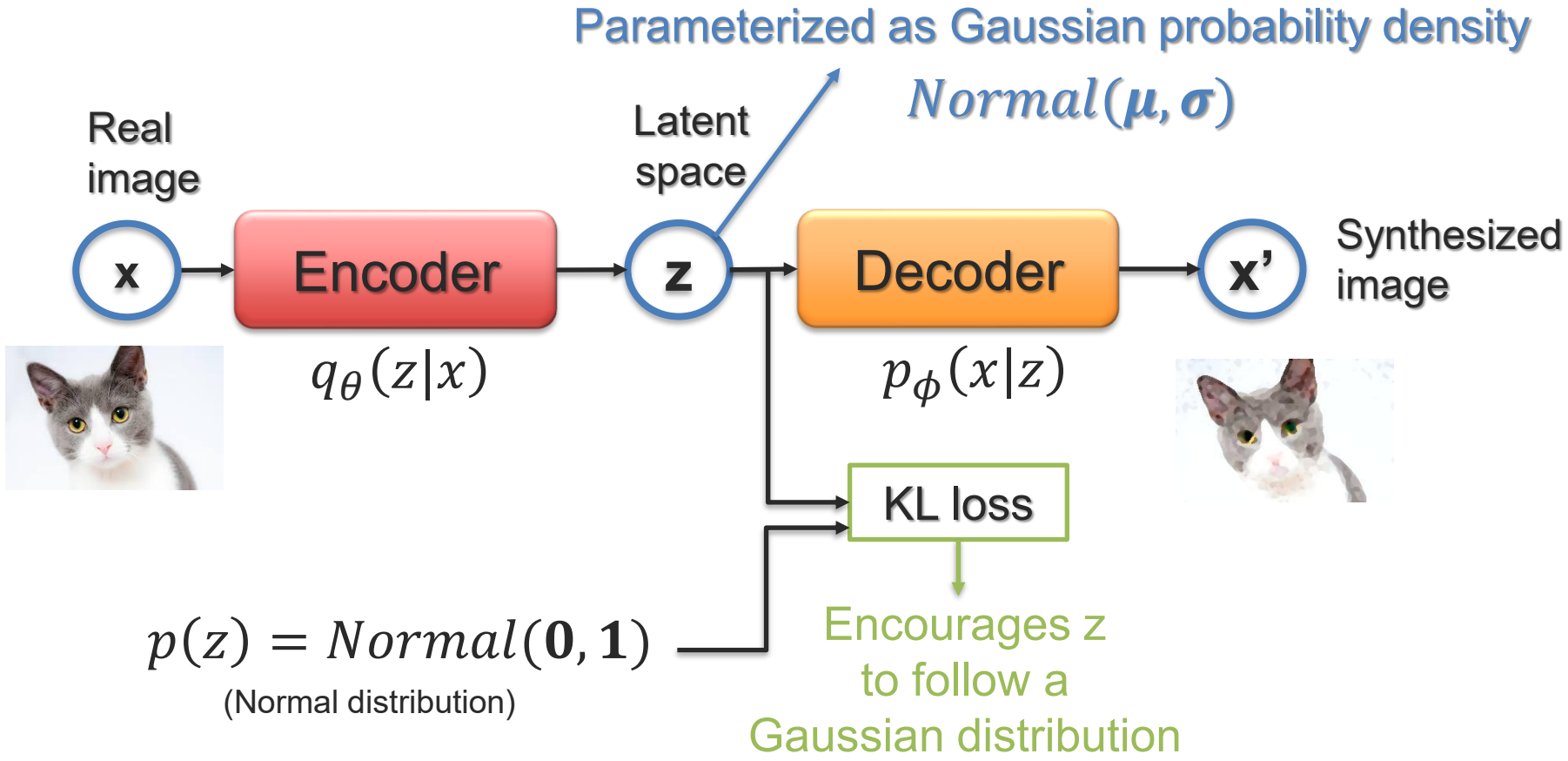
Auto-encoder



After learning this autoencoder,
can I input any z vector in the decoder?

[Kingma & Welling, Auto-Encoding Variational Bayes. ICLR 2014]

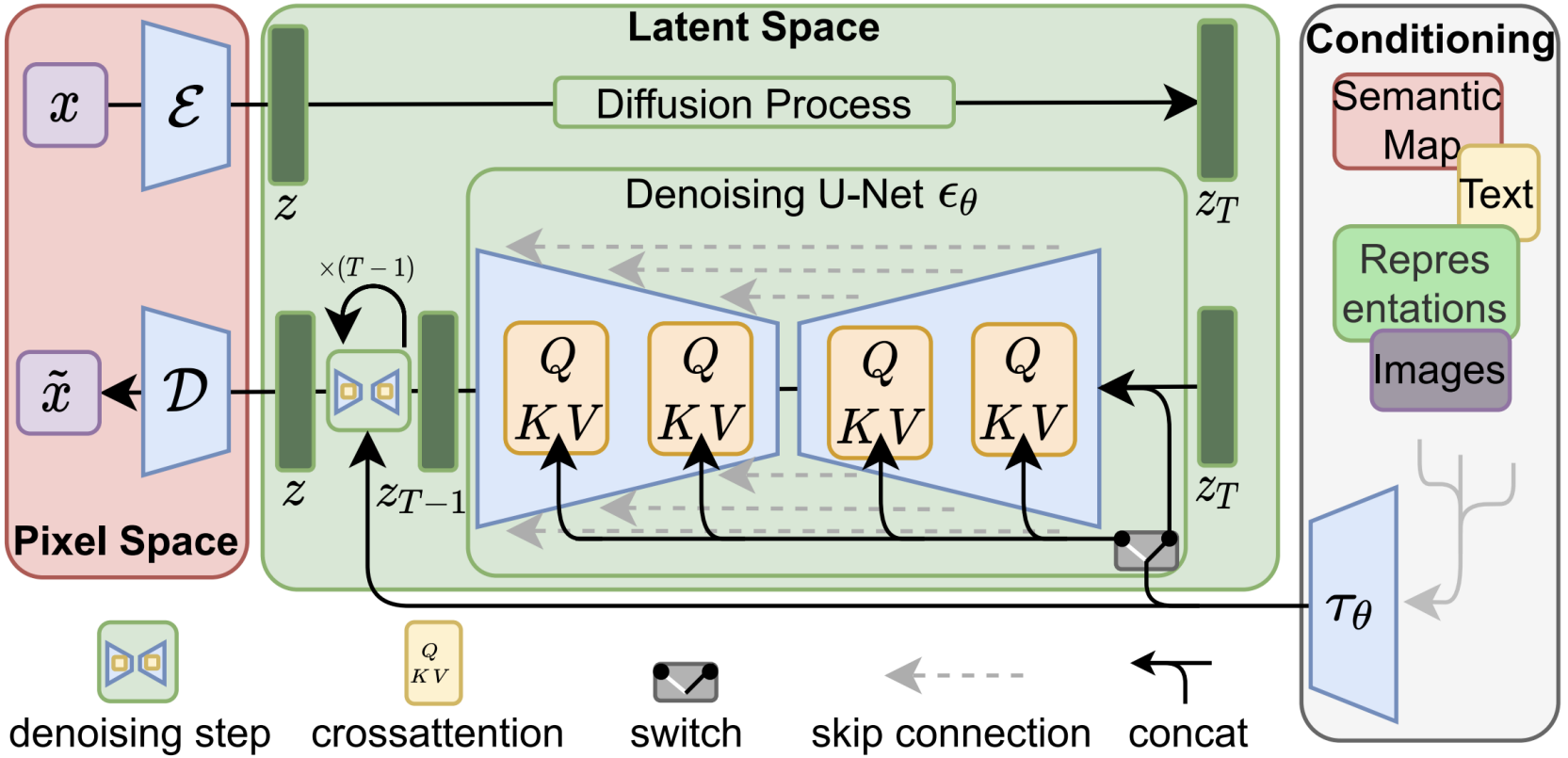
Variational Autoencoder



[Kingma & Welling, Auto-Encoding Variational Bayes. ICLR 2014]

Latent Diffusion (Architecture behind Stable Diffusion v1)

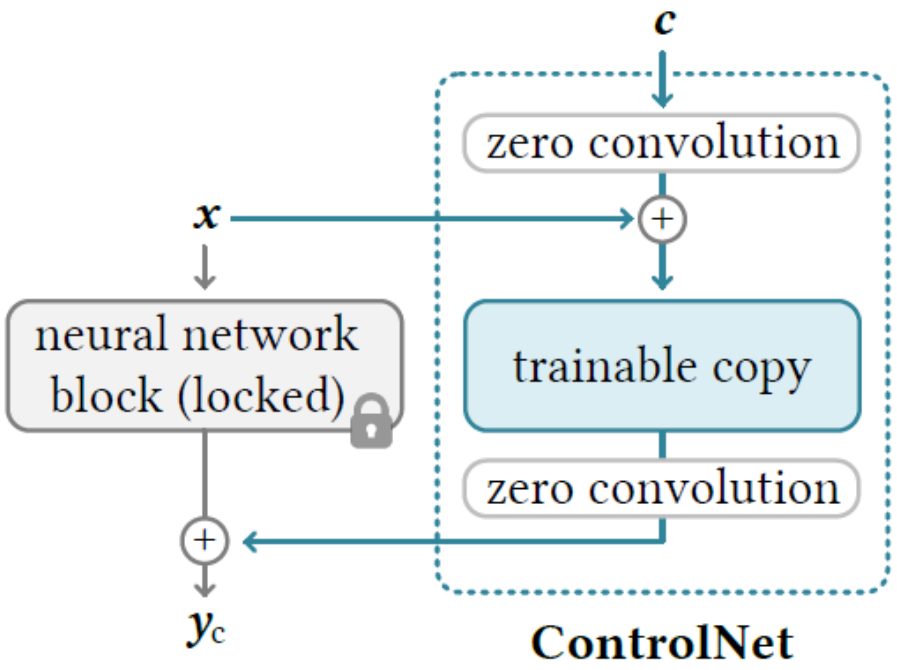
Key Idea: Don't denoise pixels. Denoise a compressed, **continuous** latent — an autoencoder (encoder and decoder) handles the pixels.

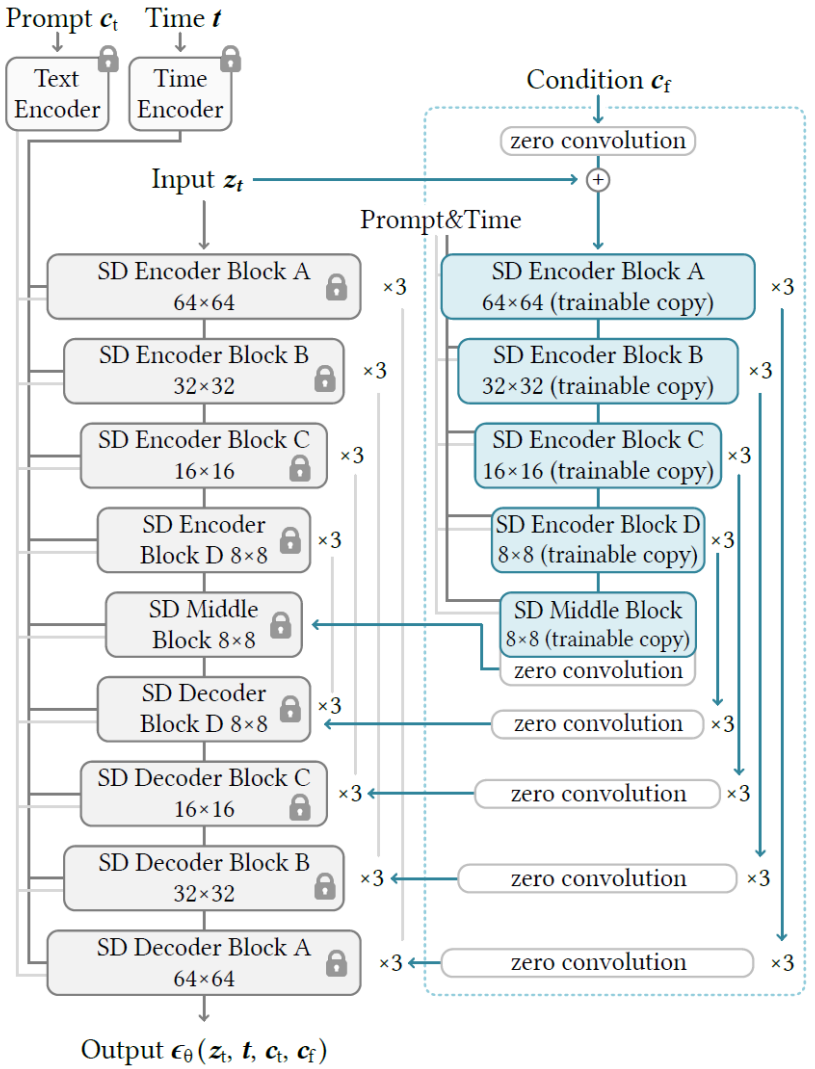


Text conditioning: Cross-attention into the denoising network, at every resolution, at every step

[Rombach, Blattmann, Lorenz, Esser & Ommer, High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022]

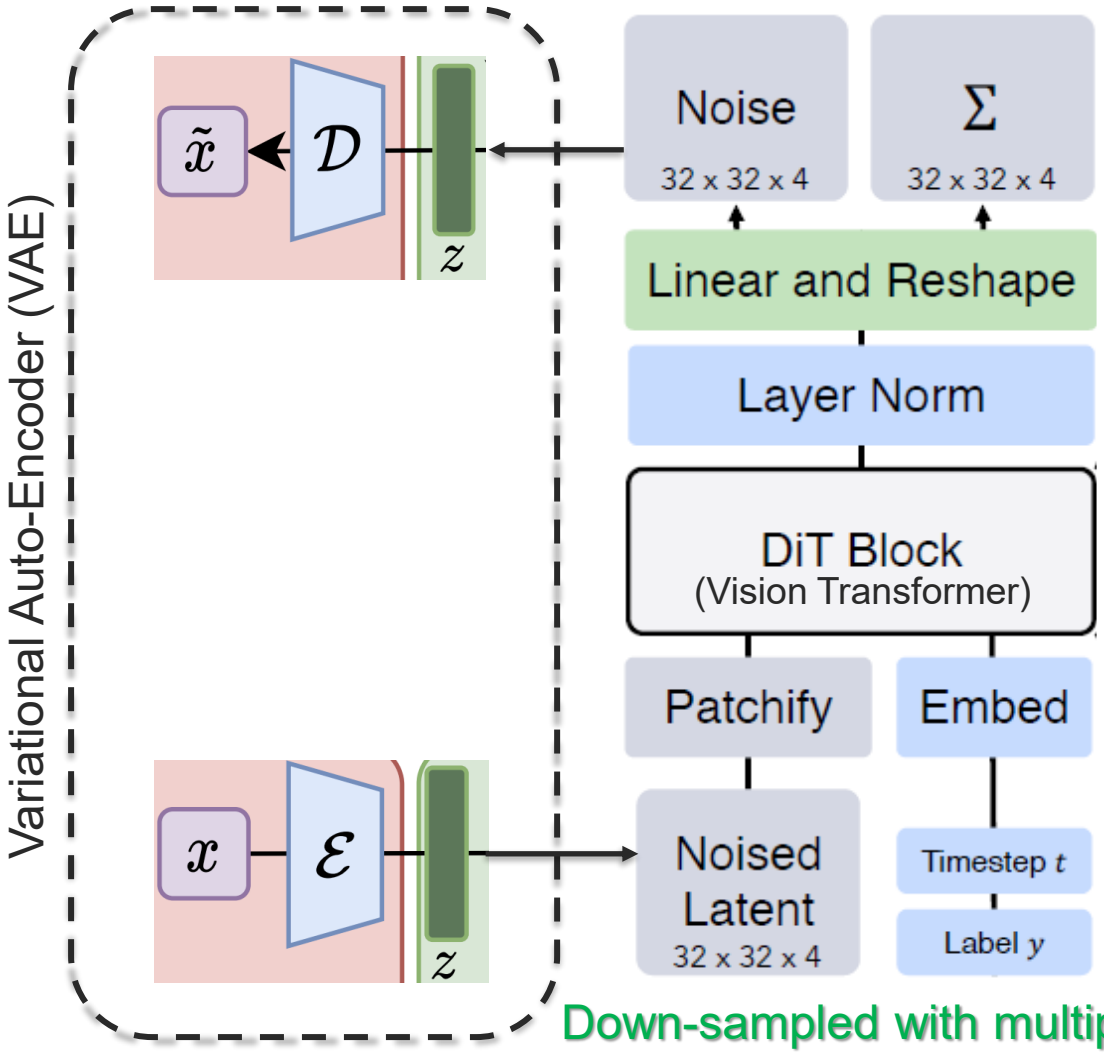
ControlNet: Adding New Conditioning to a Pretrained Diffusion Model





[Zhang, Rao & Agrawala, Adding Conditional Control to Text-to-Image Diffusion Models. ICCV 2023]

Diffusion Transformer



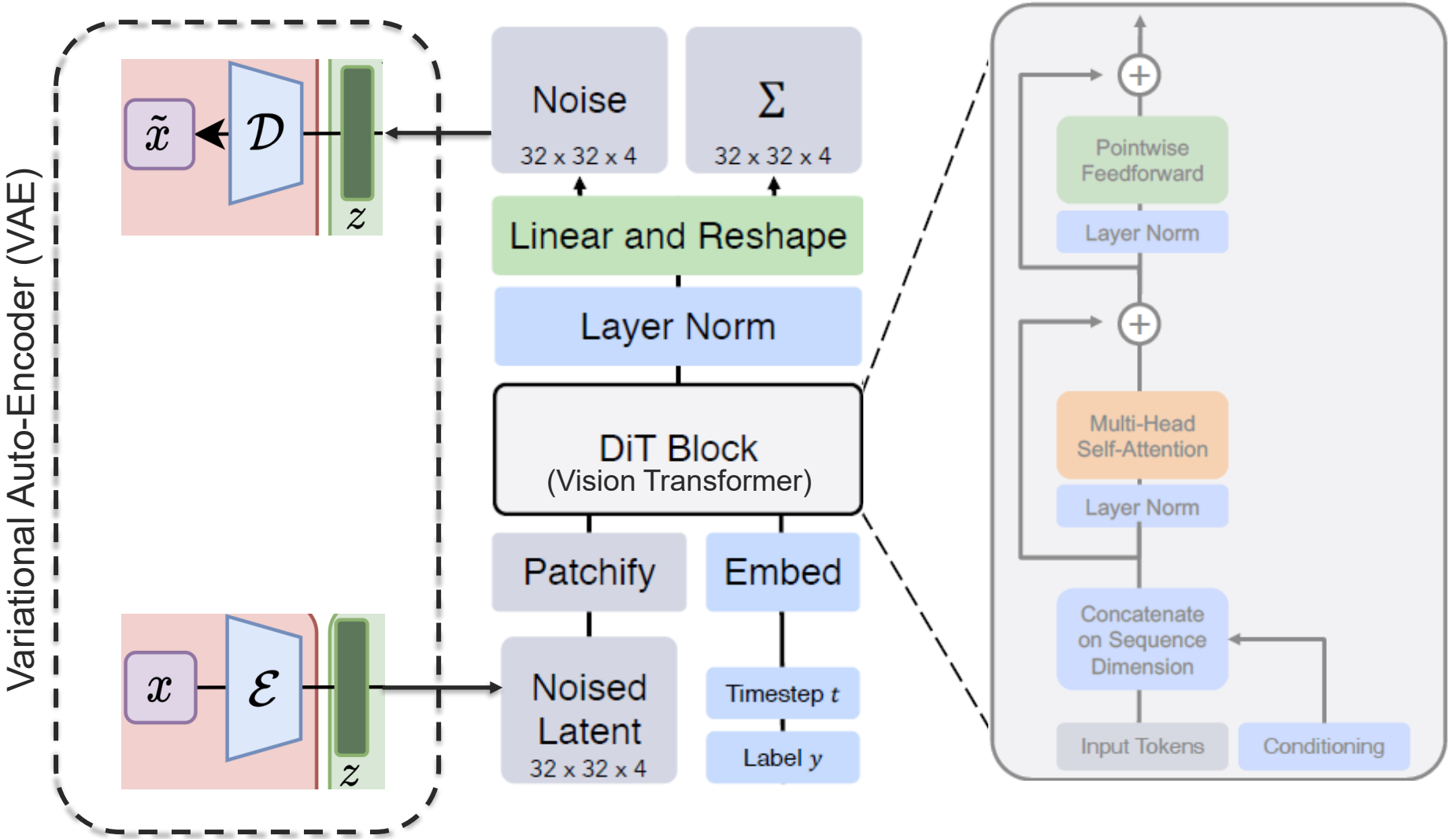
Also predict variance of the Gaussian (for noise model)

Object label is the conditioning signal

Down-sampled with multiple channels

[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

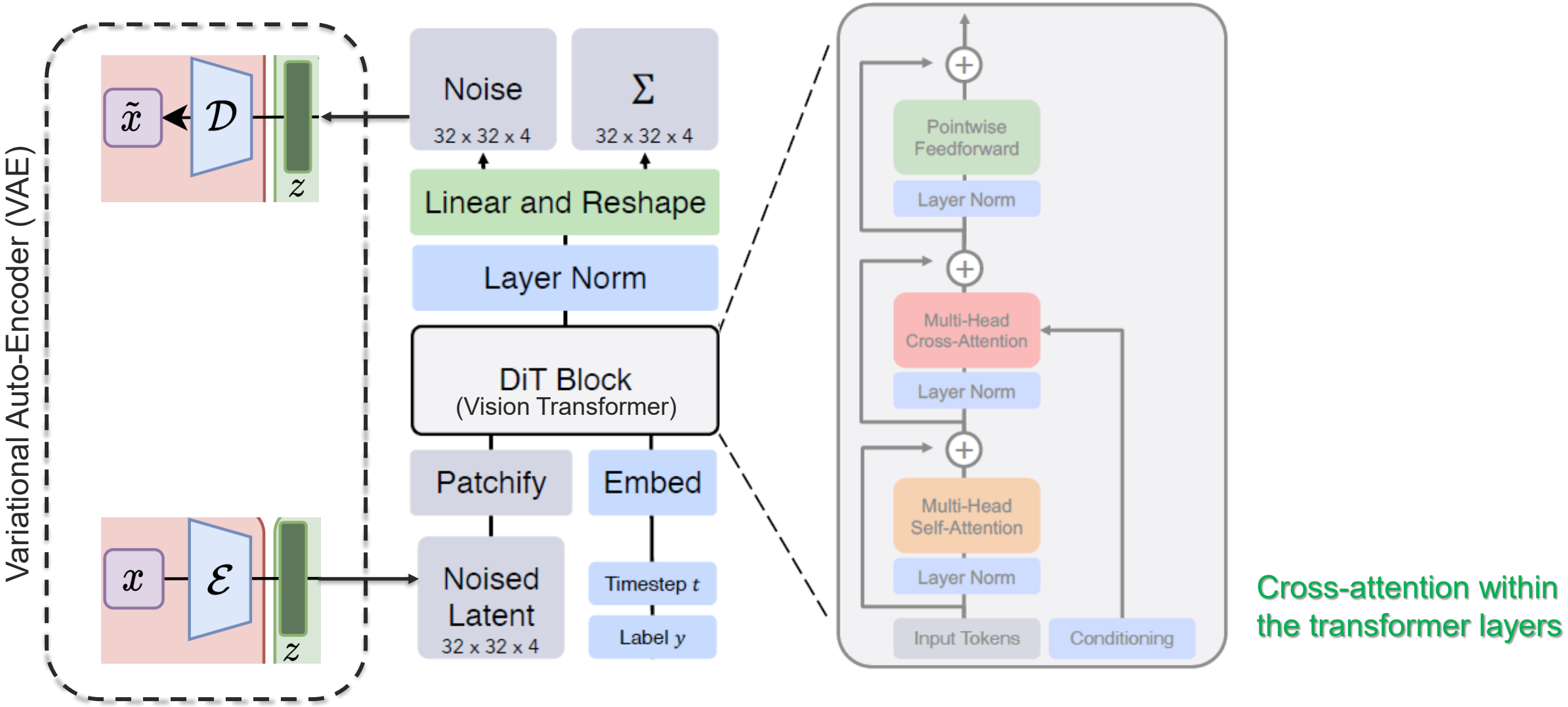
Conditioning Strategy A: In-Context Conditioning



Conditioning simply added as extra tokens

[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

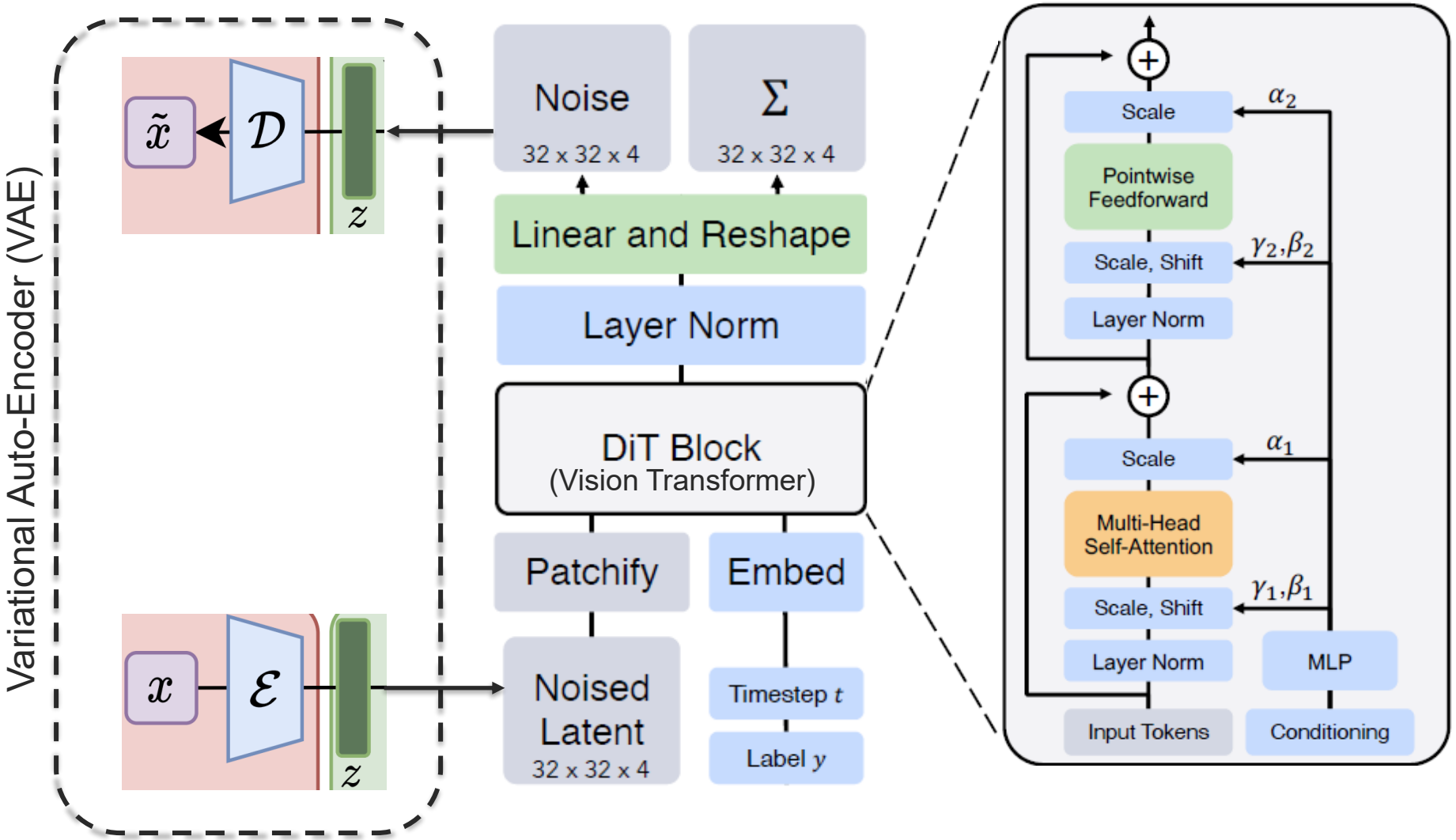
Conditioning Strategy B: Cross-Attention



Cross-attention within the transformer layers

[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

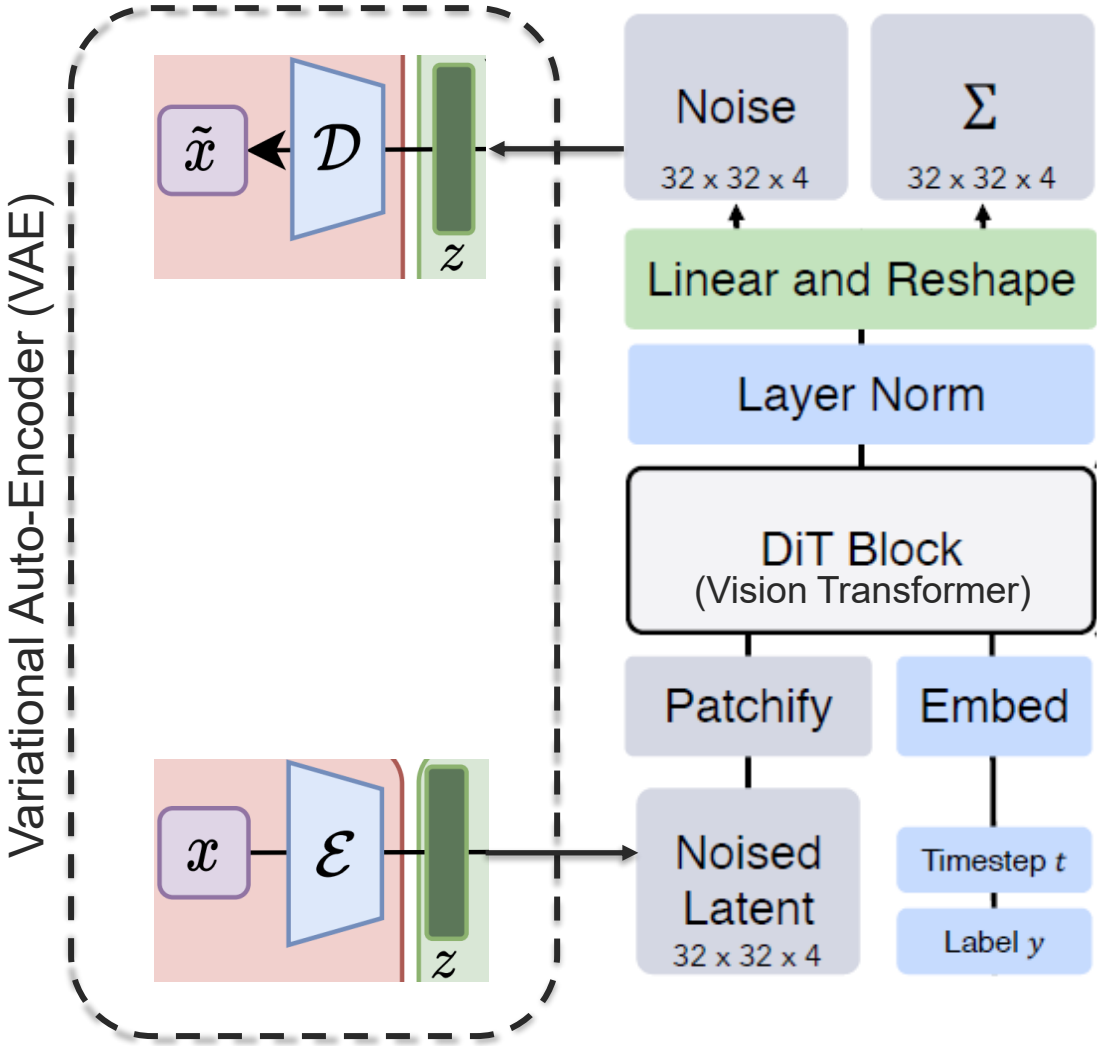
Conditioning Strategy C: Adaptive Layer Normalization (AdaLN)

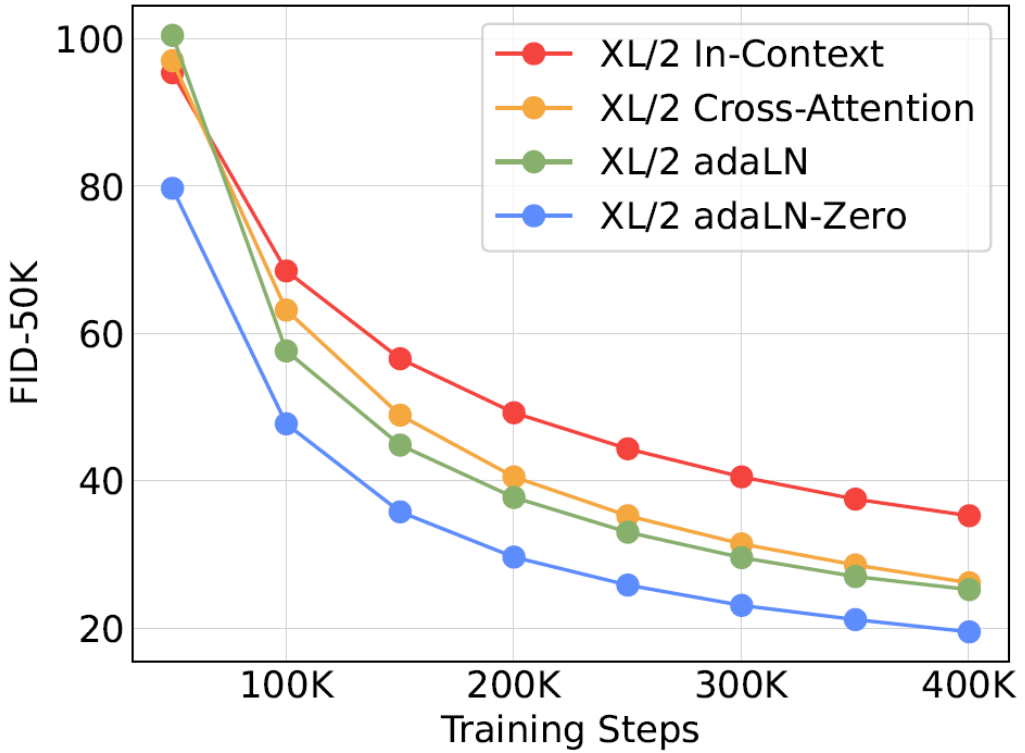


Global scaling and shift: Conditioning is not applied differently to different visual tokens

[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

Comparing Conditioning Strategies





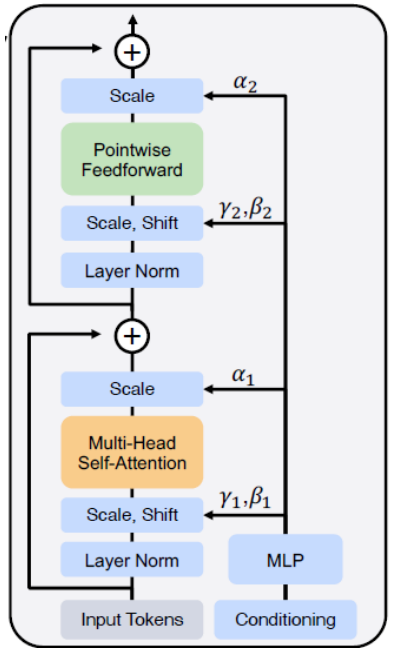
"Zero": start each block as the identity instead of random initialization

[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

Multimodal Diffusion Transformer (MM-DiT) – aka Stable Diffusion 3

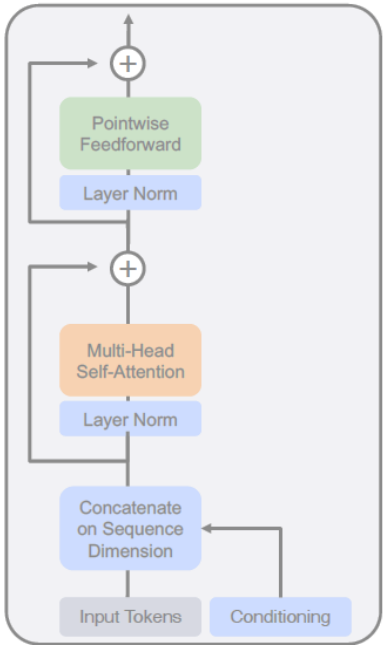
Strategy C

Global adaptation (AdaLN)



Strategy A

In-Context Conditioning



Two streams, separate weights, joined with single shared self-attention

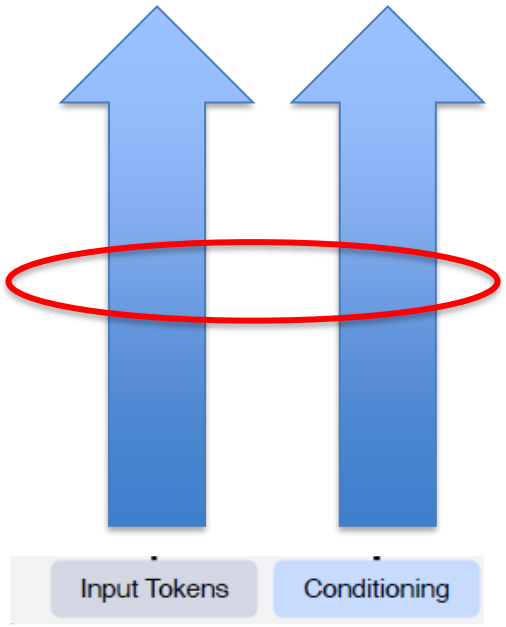
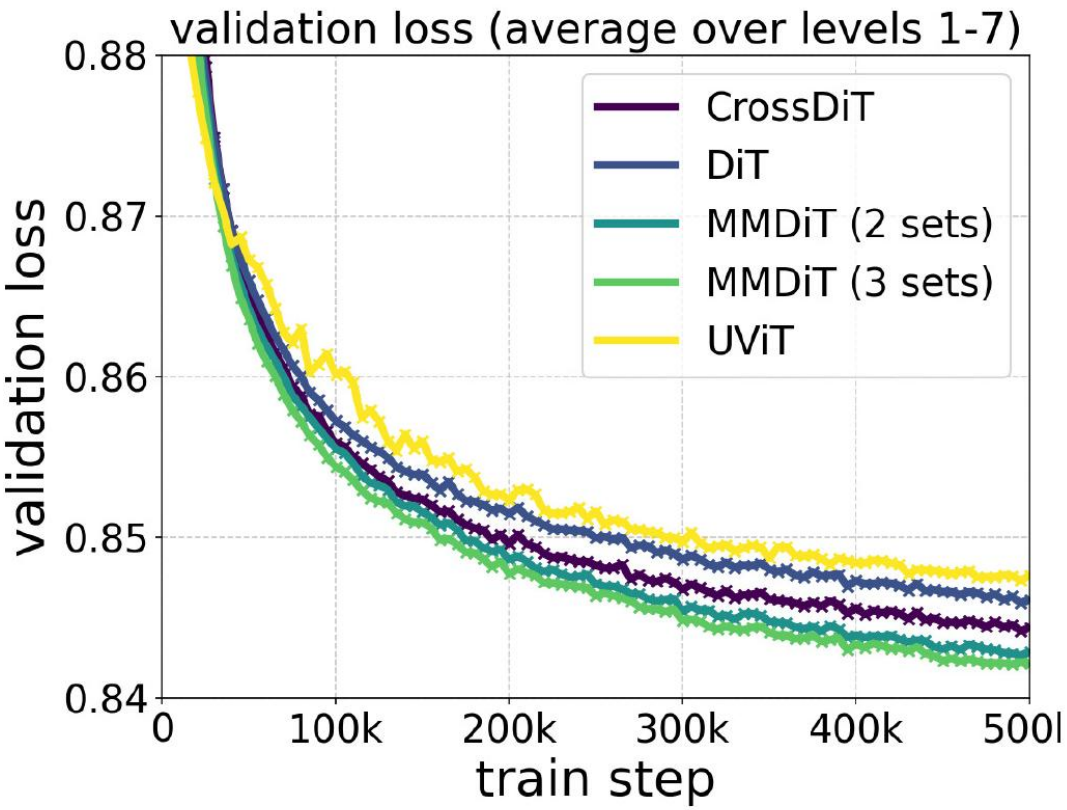


Image tokens and text tokens each have their own weights. Only the attention is shared.

[Esser et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. ICML 2024]

Two Separate Streams was Important



Two streams, separate weights, joined with single shared self-attention

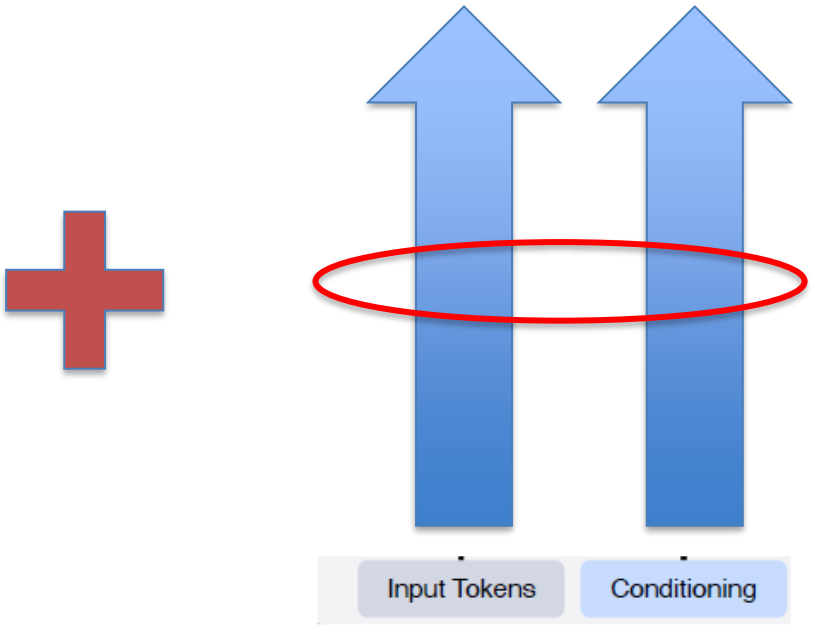


Image tokens and text tokens each have their own weights. Only the attention is shared.

[Esser et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. ICML 2024]

Classifier-Free Guidance

At training time:

$$c \leftarrow \emptyset \quad \text{with probability} \approx 0.1$$

At inference time:

Two forward passes: $\varepsilon_\theta(z_t, t, c)$ and $\varepsilon_\theta(z_t, t, \emptyset)$

and then combined: $\tilde{\varepsilon} = \varepsilon_\theta(z_t, t, \emptyset) + w\big(\varepsilon_\theta(z_t, t, c) - \varepsilon_\theta(z_t, t, \emptyset)\big)$

guidance scale

[Ho & Salimans, Classifier-Free Diffusion Guidance. arXiv 2022]

Where Does the Condition Enter?

Where	How	Where you saw it
The input	extra tokens in the sequence	Strategy A — MM-DiT
The hidden layers	cross-attention	Strategy B — Latent Diffusion
	scale & shift	Strategy C (AdaLN) — DiT, MM-DiT
	a side network adds features	ControlNet
The output	conditional - unconditional, scaled by w	Classifier-free guidance

Note: Also possible: edit z_t between steps, with no training at all (Attend-and-Excite, 2023)