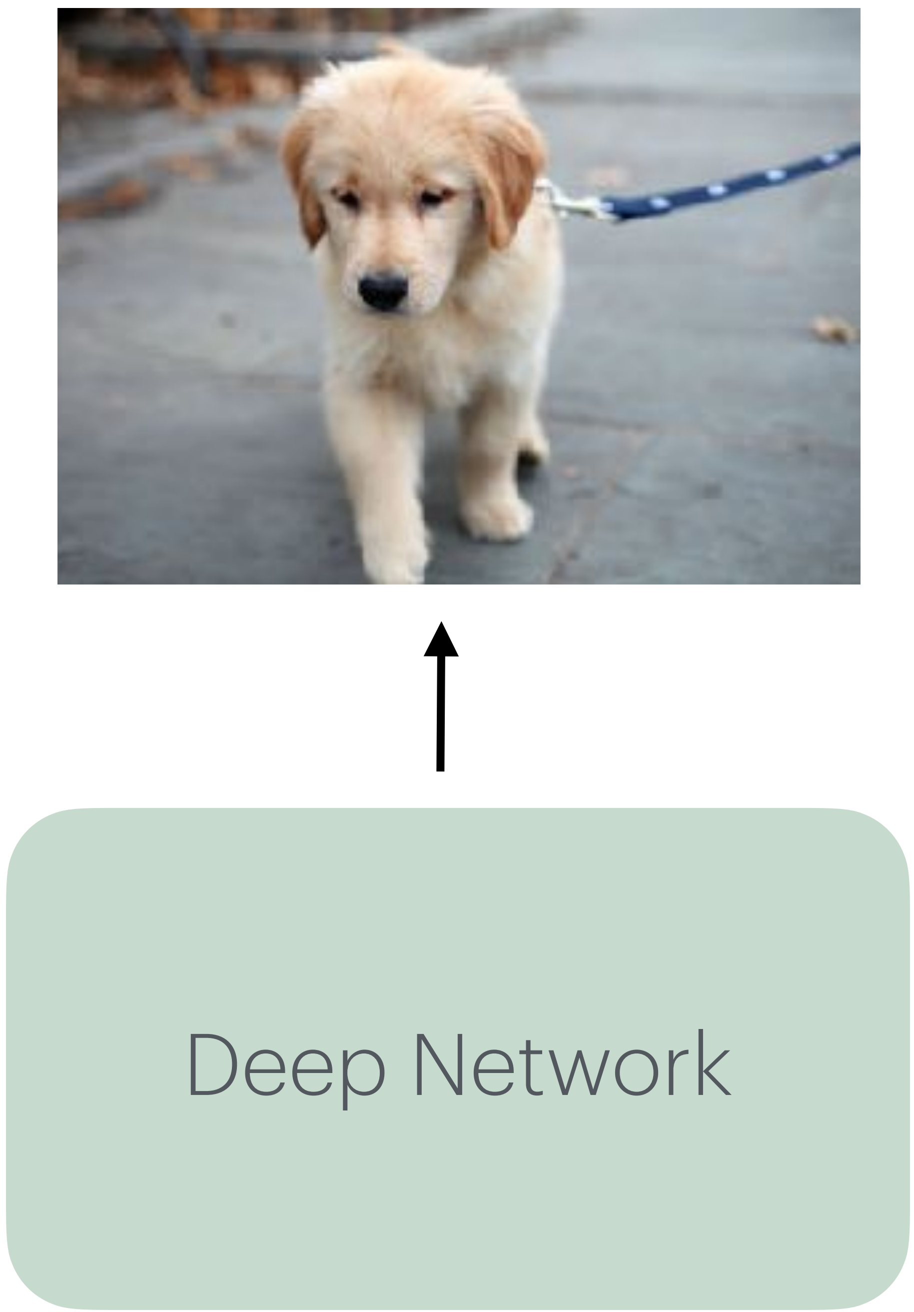
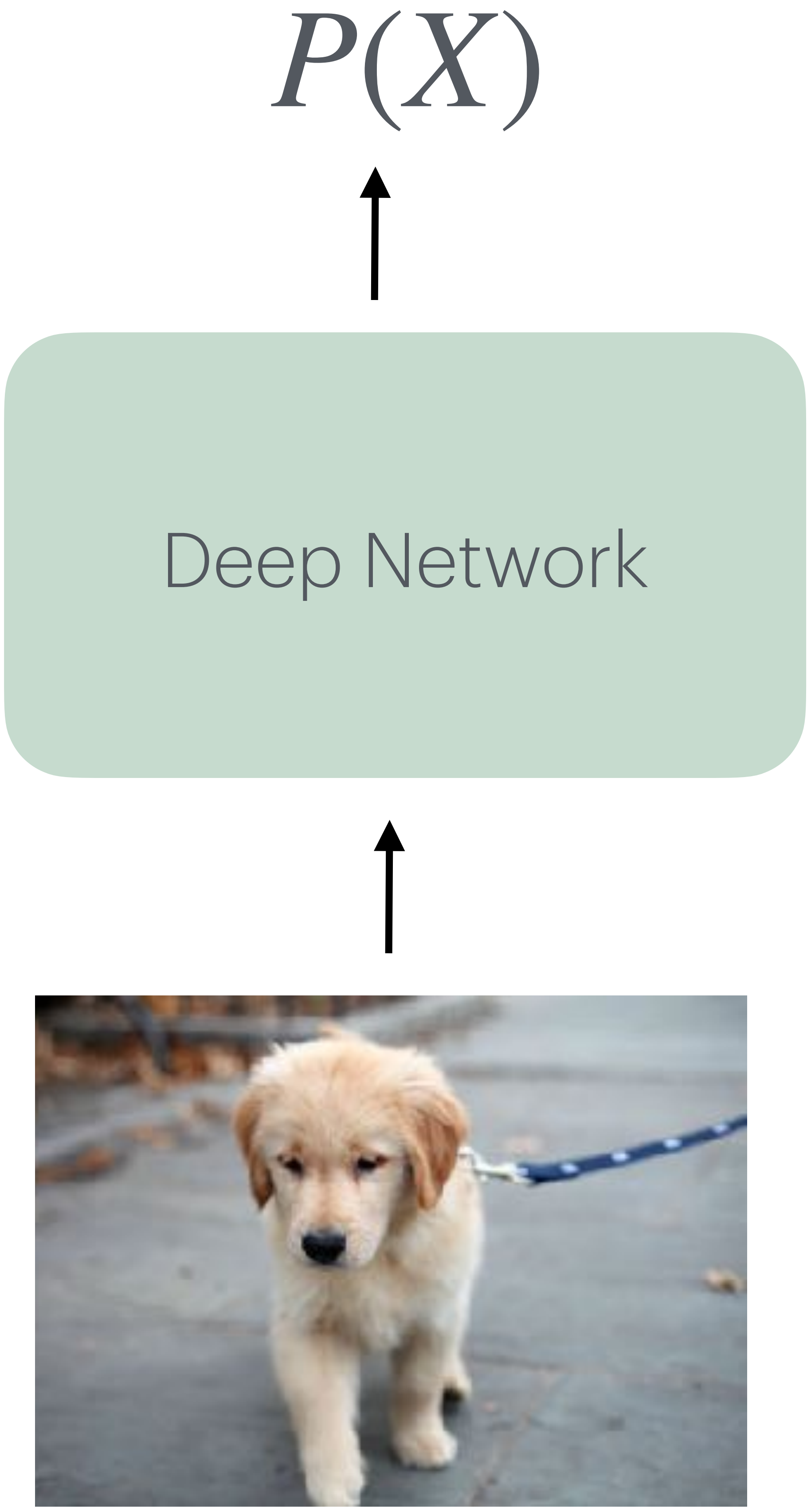
Latent Diffusion and State-of-the-Art Models

Philipp Krähenbühl, UT Austin

Generative models

- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$





Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x|z)$ or $f : z \to x$

z Deep Network 

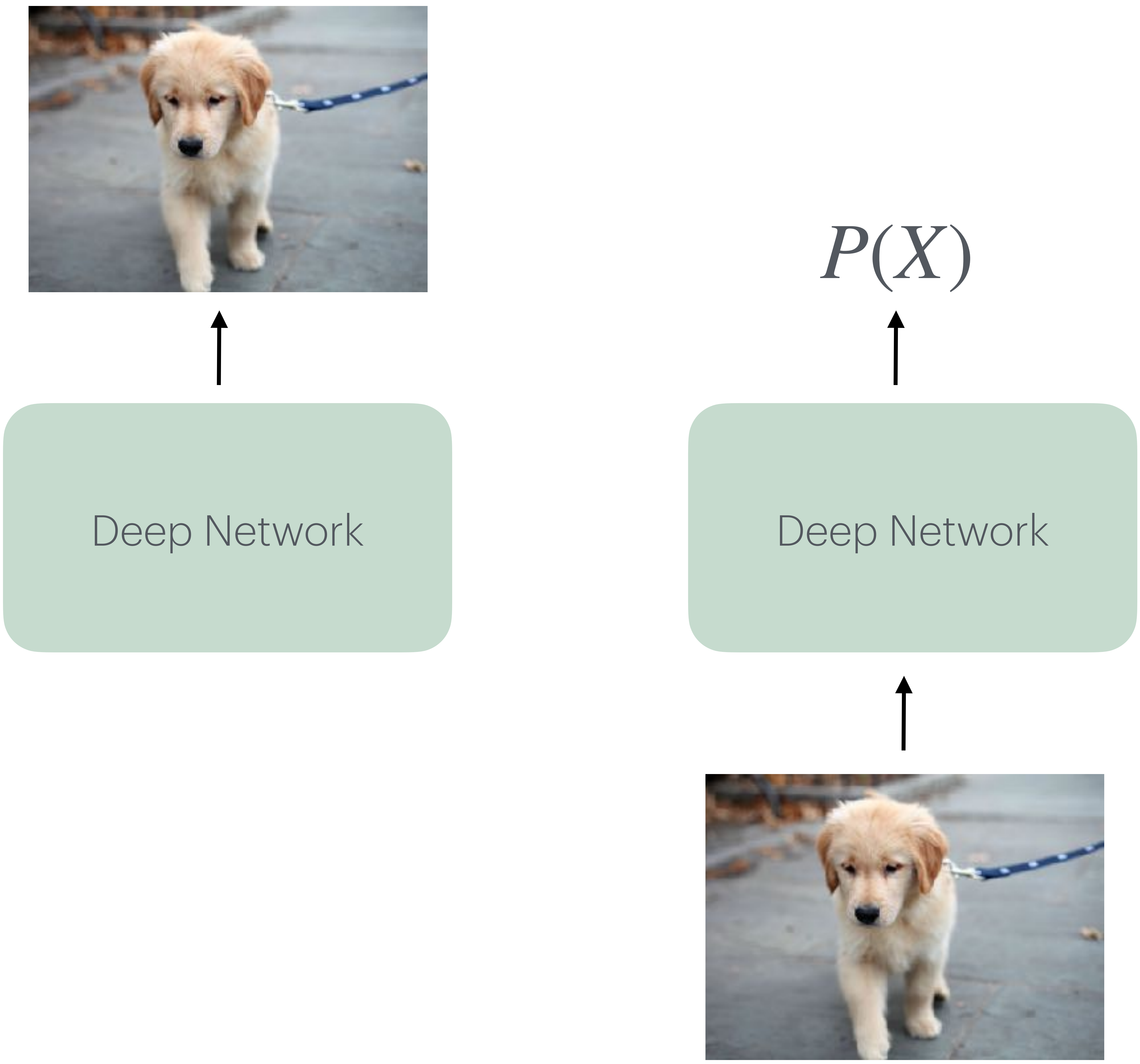
Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

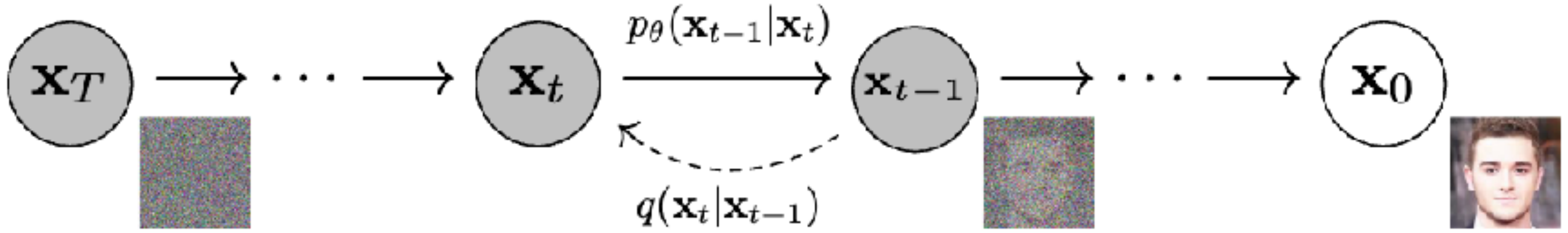
 Deep Network $P(X)$

Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all x
  - Impossible to compute (in general)



Diffusion Process



Algorithm 1 Training

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

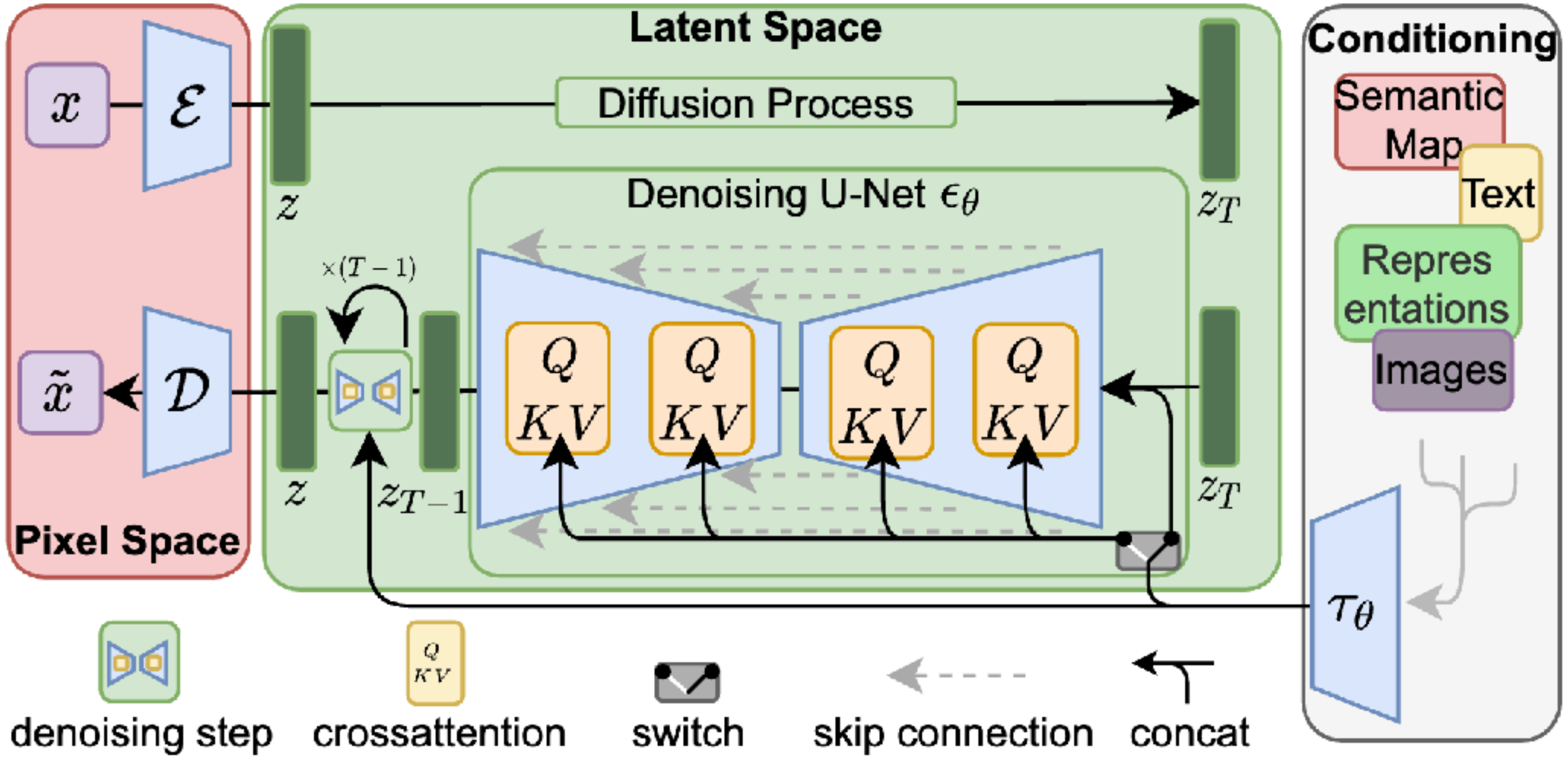
[1] Denoising Diffusion Probabilistic Models. Jonathan Ho, et al. 2020.
[2] Generative Modeling by Estimating Gradients of the Data Distribution. Yang Song, et al. 2019

Diffusion

- Very good image quality
- Not easily controllable
- Computationally quite expensive
    - Multiple sampling steps
    - Fairly high resolution inputs and outputs required (original image size)

Latent Diffusion

- Auto-encoder + Diffusion
  - Similar to VQVAE + Auto-regressive
- Speeds up training and generation
  - Lower resolution diffusion
  - Auto-encoders are fast
- Higher resolution outputs



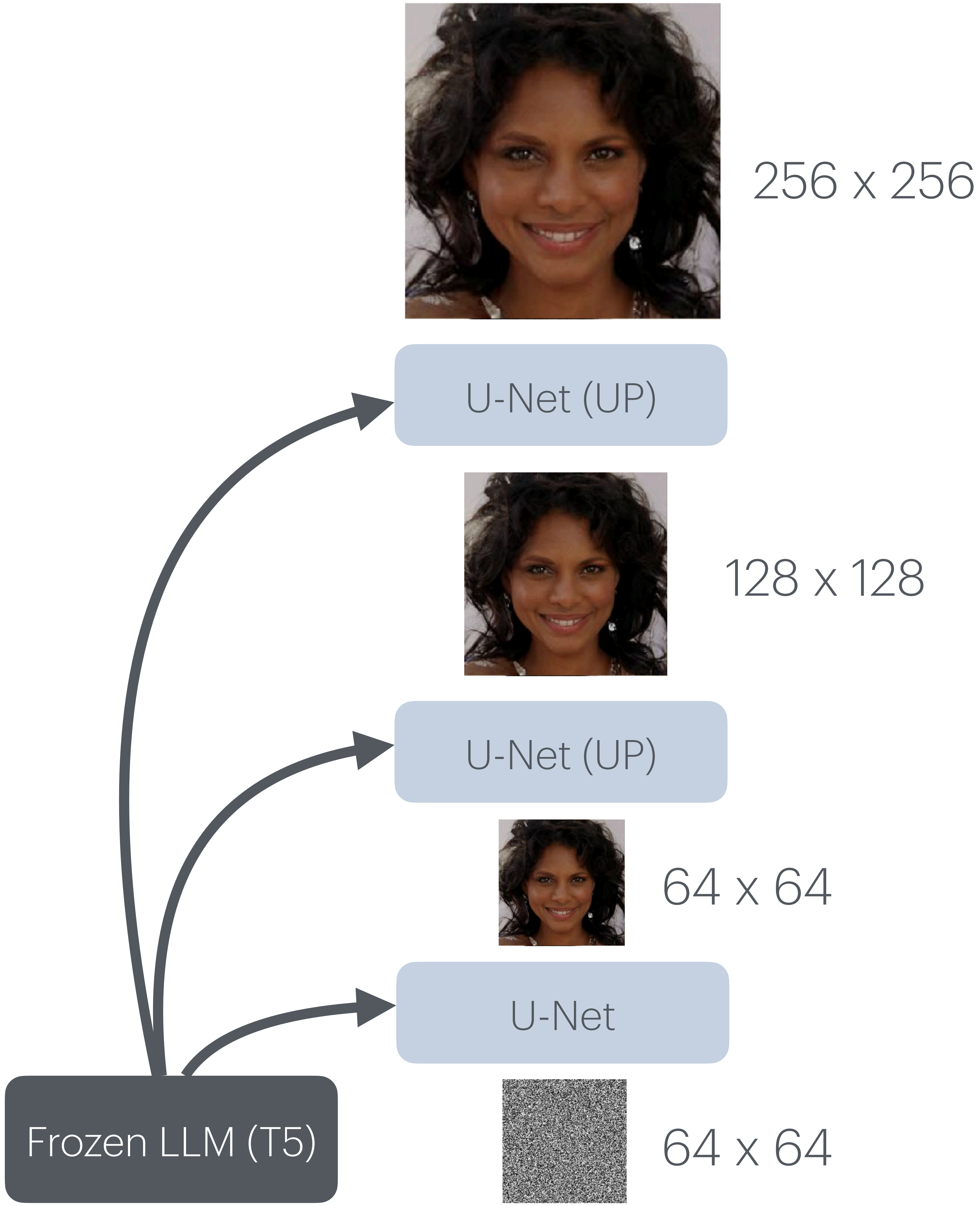
[3] High-Resolution Image Synthesis with Latent Diffusion Models. Robin Rombach, et al. 2021.

Latent Diffusion

[3] High-Resolution Image Synthesis with Latent Diffusion Models. Robin Rombach,, et al. 2021.

Imagen

- First really large scale diffusion model
  - 800M+ image-text pairs
- Frozen LLM
- Lower resolution diffusion 64x64
  - Upsampling to 1024



[4] Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Chitwan Saharia, et al. 2022.

Imagen

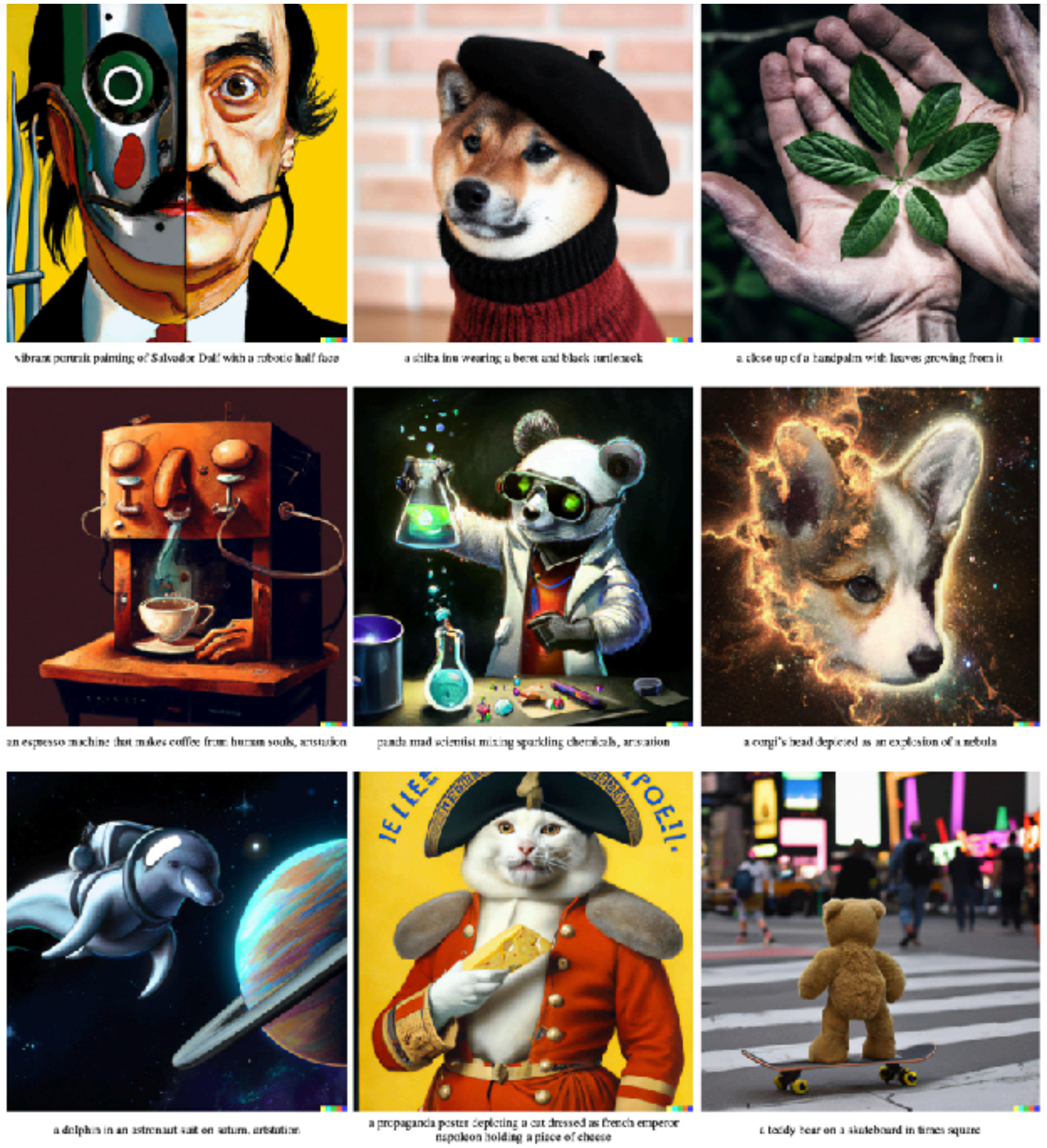
Results

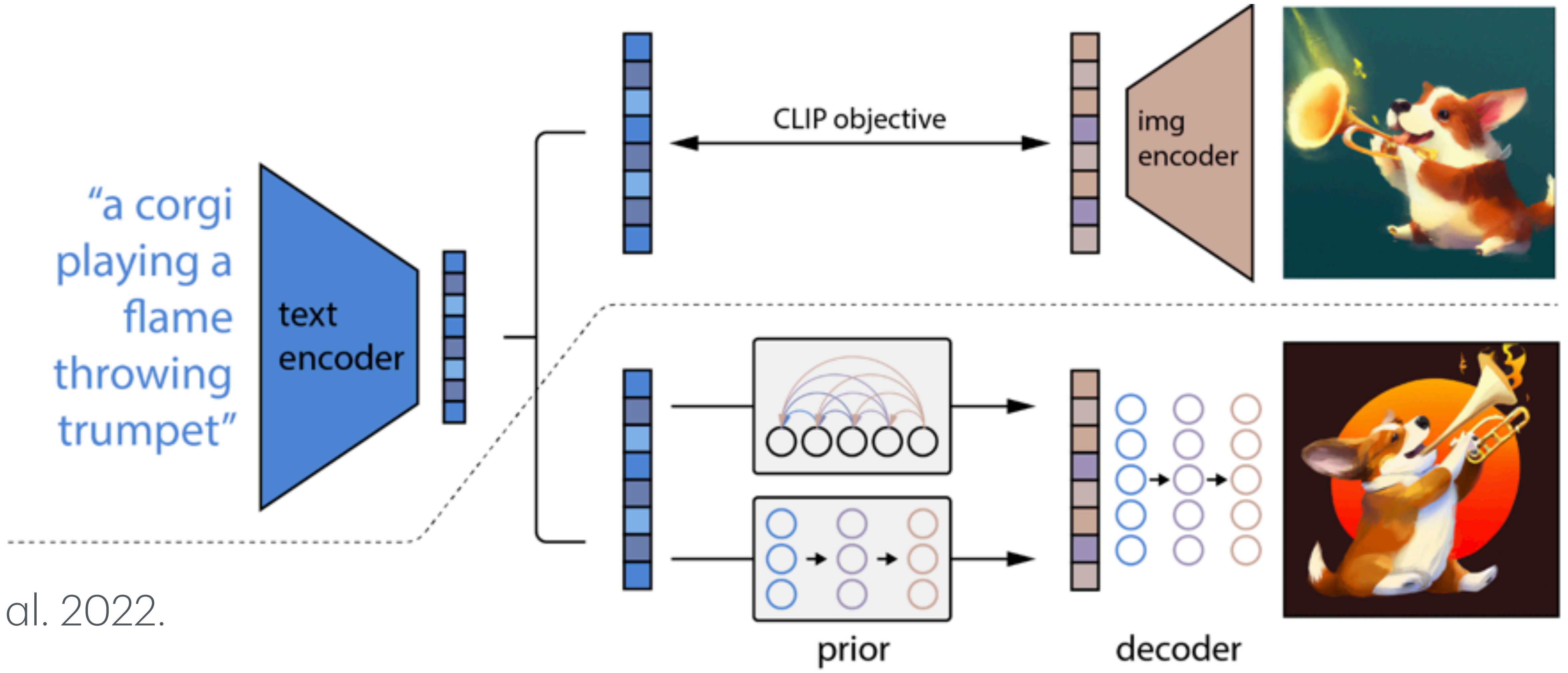
A chrome-plated duck with a golden beak arguing with an angry turtle in a forest

The Toronto skyline with Google brain logo written in fireworks.

DALL-E 2

- CLIP-LM conditioned diffusion
  - 64x64 results
  - Upsampling
    $$64 \times 64 \rightarrow 256 \times 256 \rightarrow 1024 \times 1024$$





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

DALL-E 3

- Better data
  - Recaptioned dataset

A fierce garden gnome warrior, clad in armor crafted from leaves and bark, brandishes a tiny sword and shield. He stands valiantly on a rock amidst a blooming garden, surrounded by colorful flowers and towering plants. A determined expression is painted on his face, ready to defend his garden kingdom.

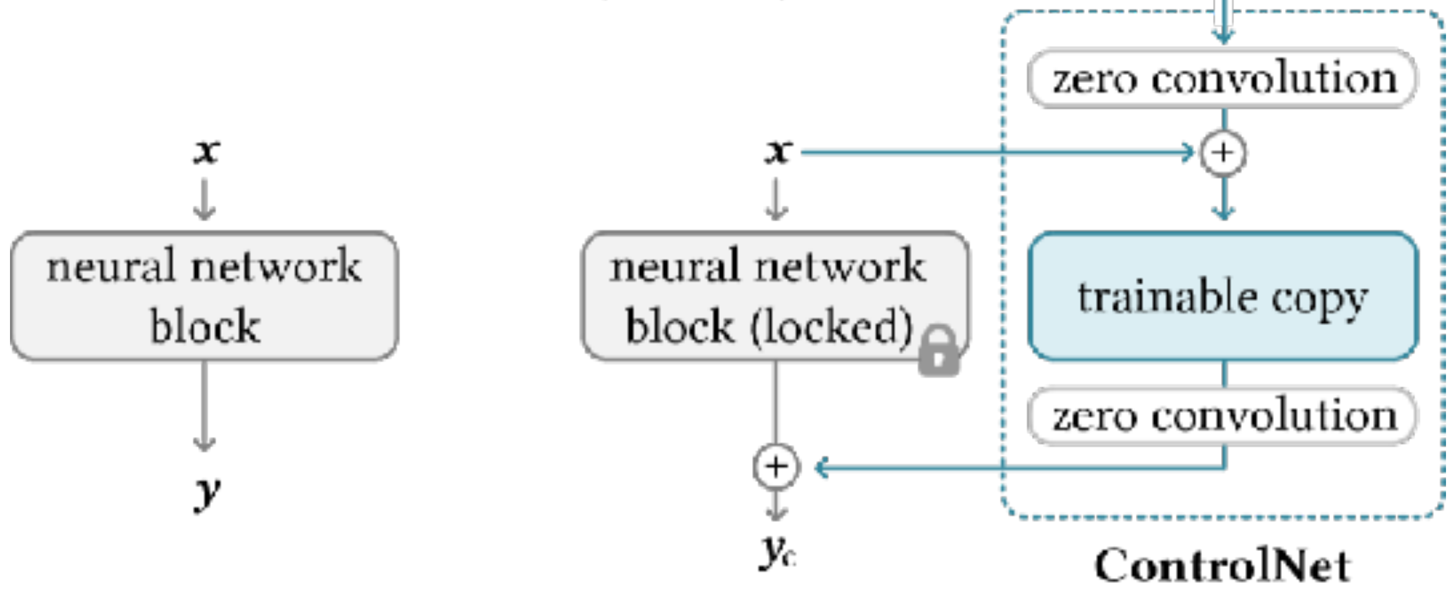
An icy landscape under a starlit sky, where a magnificent frozen waterfall flows over a cliff. In the center of the scene, a fire burns bright, its flames seemingly frozen in place, casting a shimmering glow on the surrounding ice and snow.

A swirling, multicolored portal emerges from the depths of an ocean of coffee, with waves of the rich liquid gently rippling outward. The portal engulfs a coffee cup, which serves as a gateway to a fantastical dimension. The surrounding digital art landscape reflects the colors of the portal, creating an alluring scene of endless possibilities.

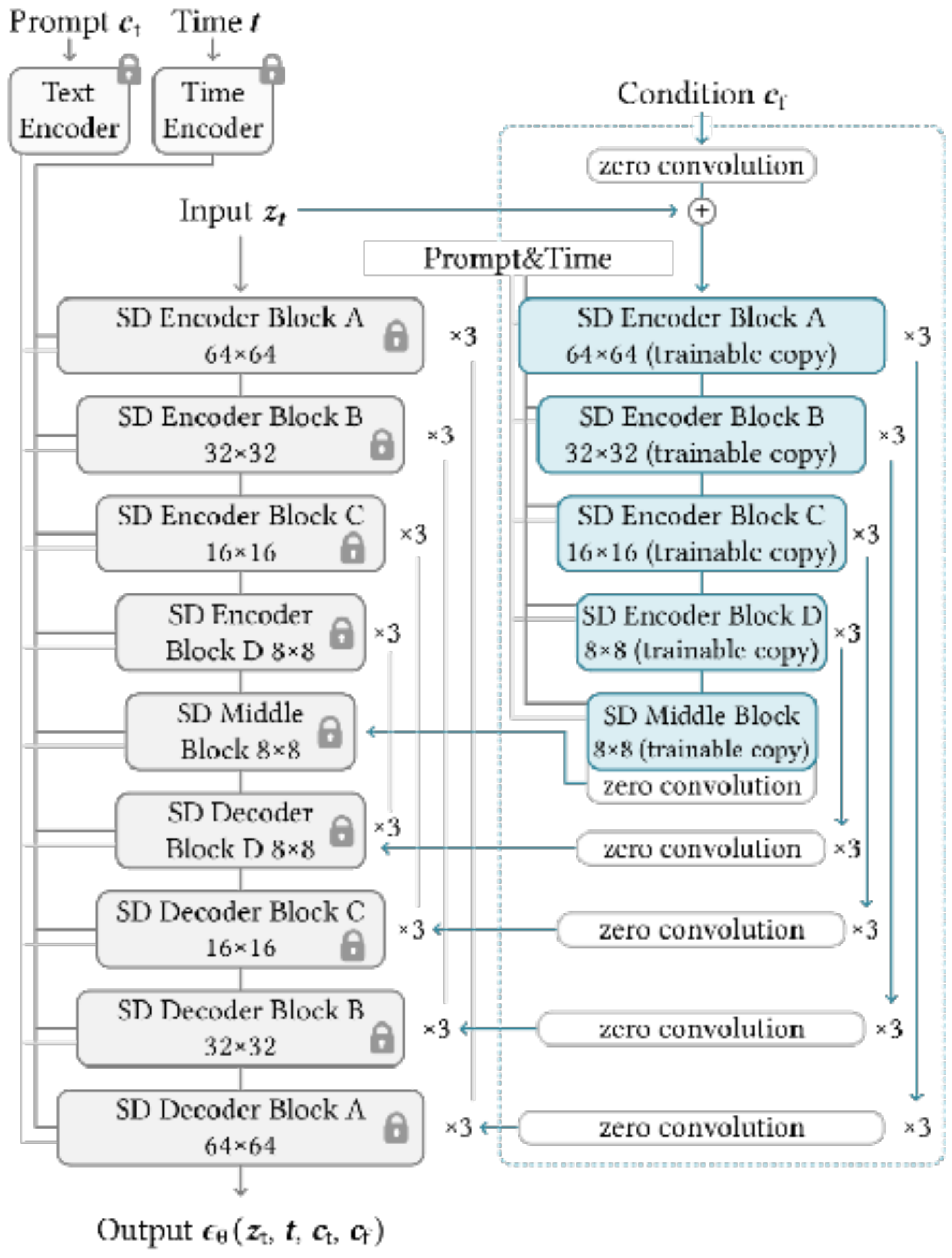
[6] Improving Image Generation with Better Captions. James Betker, et al. 2023.

ControlNet

- “Condition” on more than just text
- Start from pre-trained model
- Add copy of encoder
  - For additional input
  - Fuse with zero-initialized convolution



(a) Before (b) After

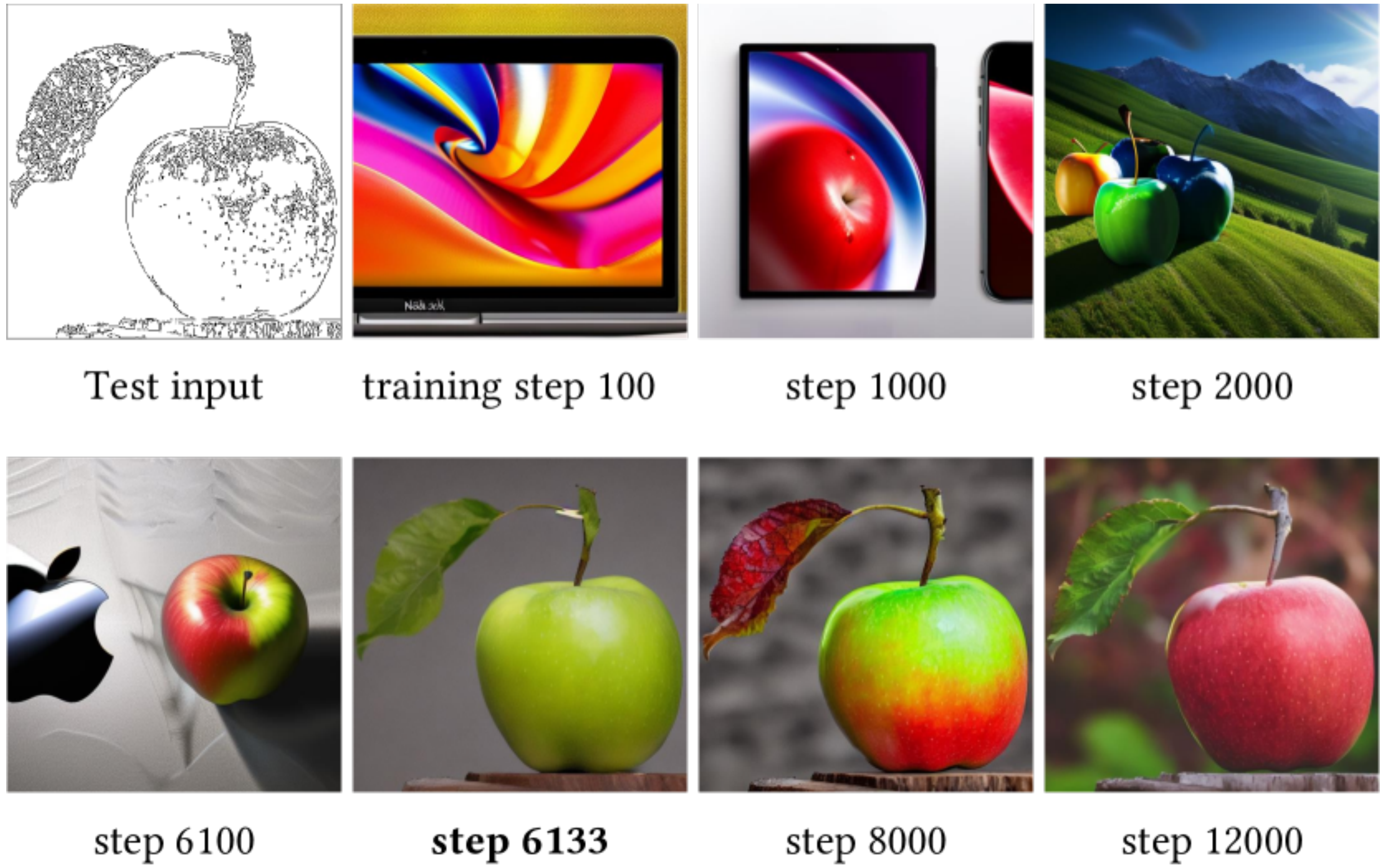


(a) Stable Diffusion (b) ControlNet

[7] Adding Conditional Control to Text-to-Image Diffusion Models. Lvmin Zhang, et al. 2023.

ControlNet

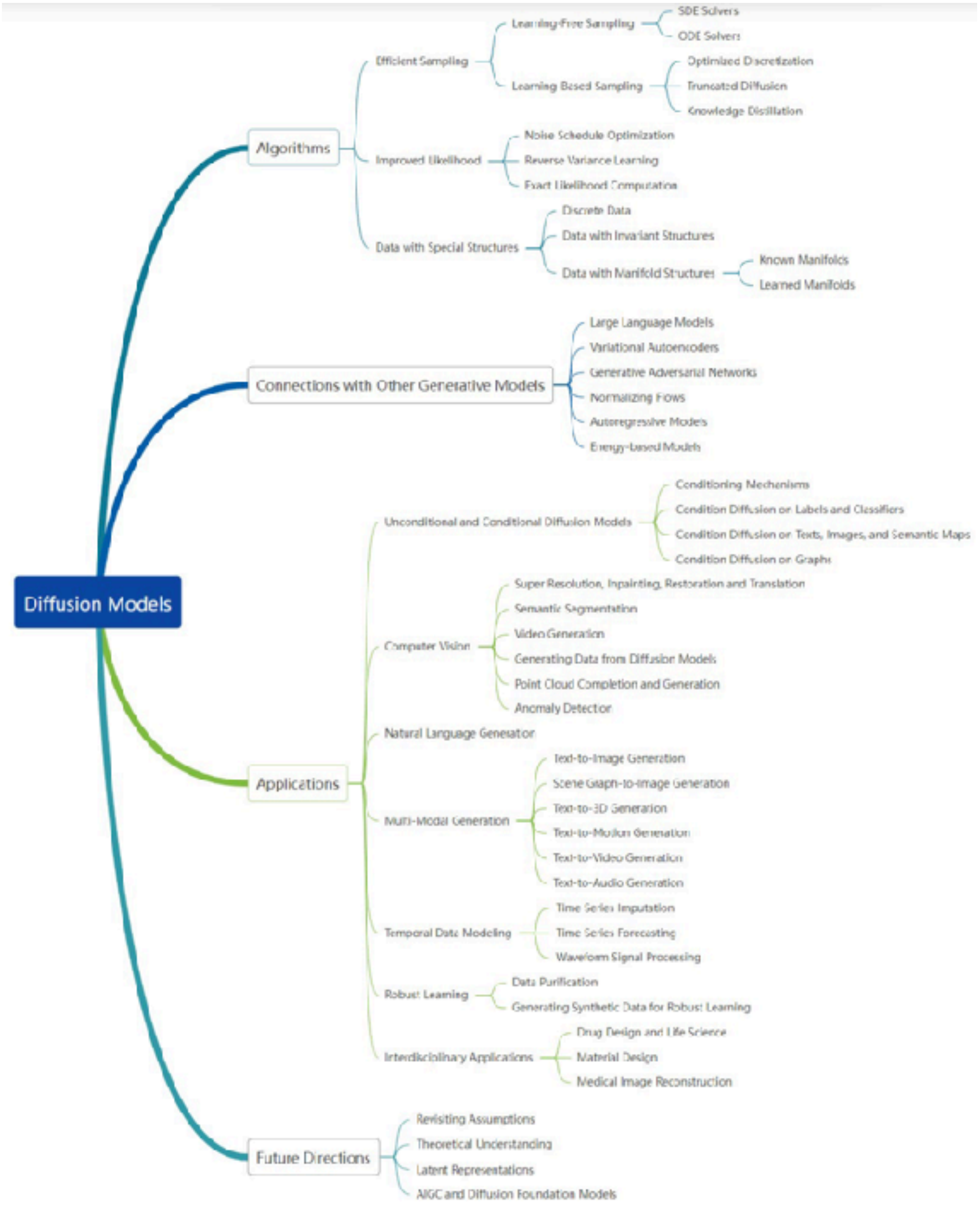
- Training objective: Denoise
  - Original image + noise
  - Continioned on auto-generated edge detections, pose tracks, ...
- Trains quite quickly



[7] Adding Conditional Control to Text-to-Image Diffusion Models. Lvmin Zhang, et al. 2023.

Diffusion is a large field

- More efficient sampling
  - One step diffusion, ...
- More efficient architectures
- More efficient training
  - Noise schedules, variance learning, ..
- ...



[8] One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. 2023.
[9] Diffusion Models: A Comprehensive Survey of Methods and Applications. Ling Yang, et al. 2022.

References

- [1] Denoising Diffusion Probabilistic Models. Jonathan Ho, et al. 2020.
- [2] Generative Modeling by Estimating Gradients of the Data Distribution. Yang Song, et al. 2019
- [3] High-Resolution Image Synthesis with Latent Diffusion Models. Robin Rombach, et al. 2021.
- [4] Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Chitwan Saharia, et al. 2022.
- [5] Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh, et al. 2022.
- [6] Improving Image Generation with Better Captions. James Betker, et al. 2023.
- [7] Adding Conditional Control to Text-to-Image Diffusion Models. Lvmin Zhang, et al. 2023.
- [8] One-step Diffusion with Distribution Matching Distillation. Tianwei Yin, et al. 2023.
- [9] Diffusion Models: A Comprehensive Survey of Methods and Applications. Ling Yang, et al. 2022.