UiT The Arctic University of Norway

NORA summer school on multi-modal learning

Multi-modal Generation

Rwiddhi Chakraborty

UiT Machine Learning Group and Visual Intelligence

Schedule Today

- 10 – 11 : Generative AI - I
- 11 – 12 : Generative AI - II
- 12 – 13: Lunch
- 13 – 14: Multi-modal Generation
- 14 – 16: Group Project

In this talk

Conditioning

Control

Edits

Cool Stuff

In this talk

Conditioning

Control

Edits

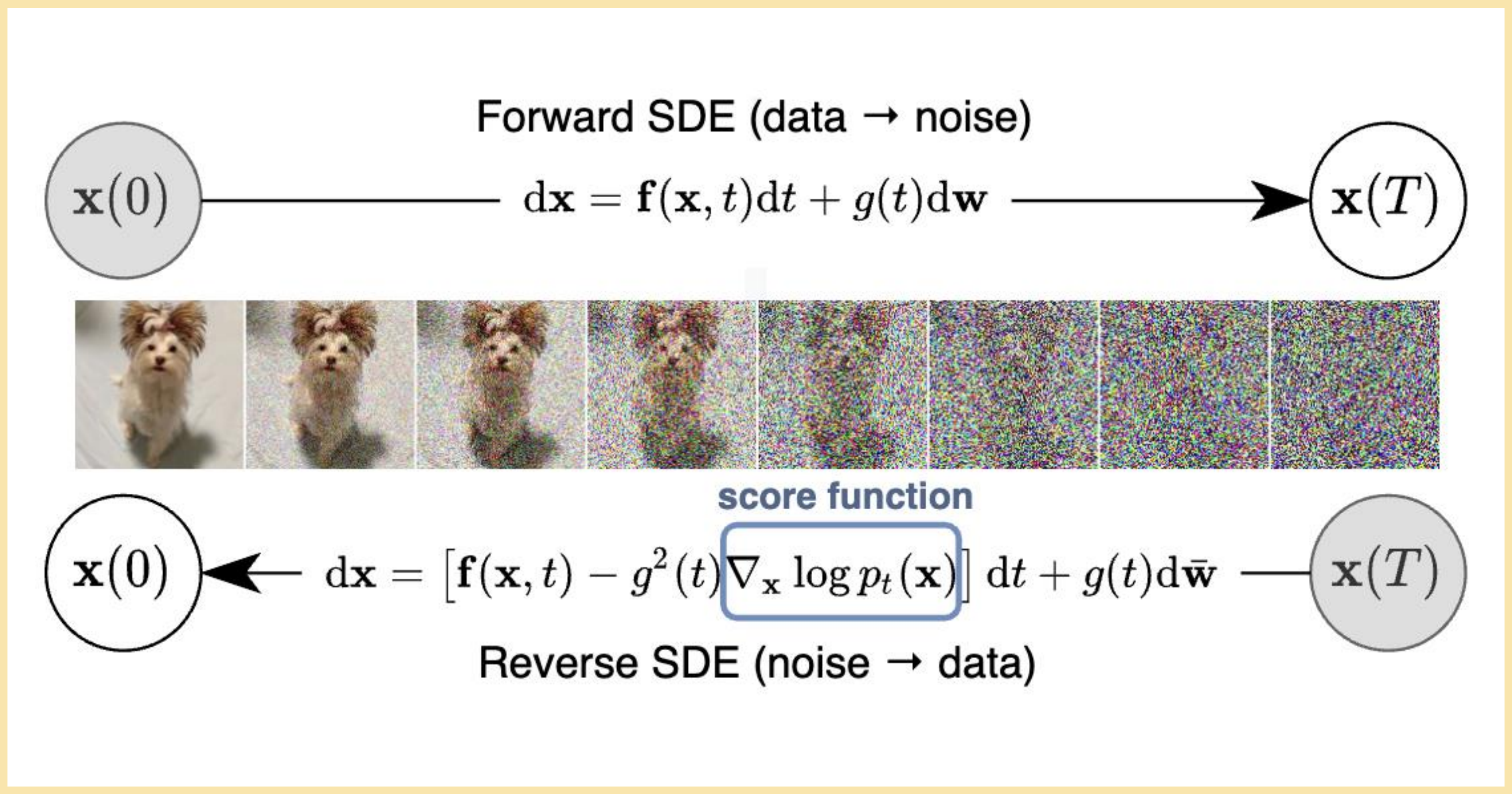
Cool Stuff

Generation: A noisy view

“Creating noise from data is easy, creating data from noise is generative modelling.”

Yang Song et al, Score-based generative modelling through stochastic differential equations.

Generation: A noisy view



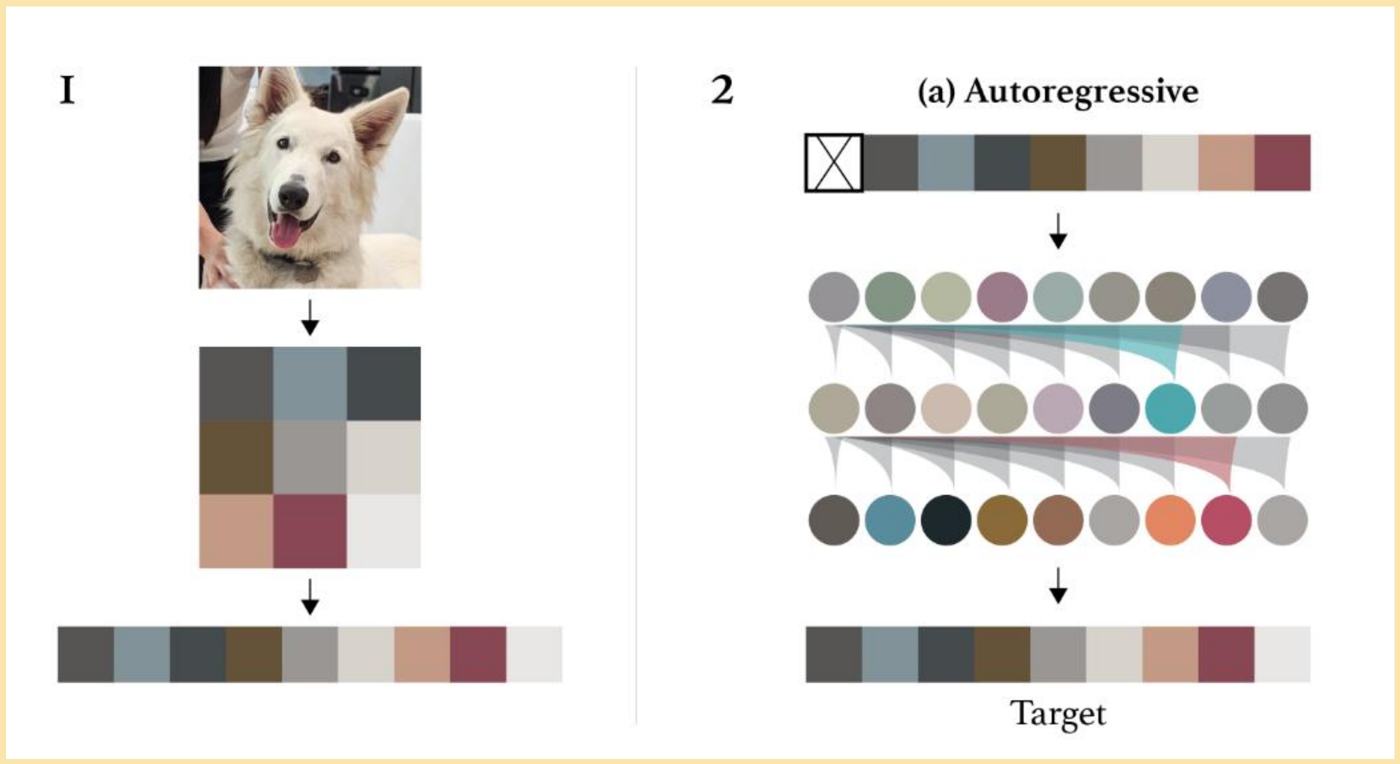
Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." arXiv preprint arXiv:2011.13456 (2020).

Generation: An autoregressive view

"….promising as our architecture uses a dense connectivity pattern which does not encode the 2D spatial structure of images yet is able to match and even outperform approaches which do…."

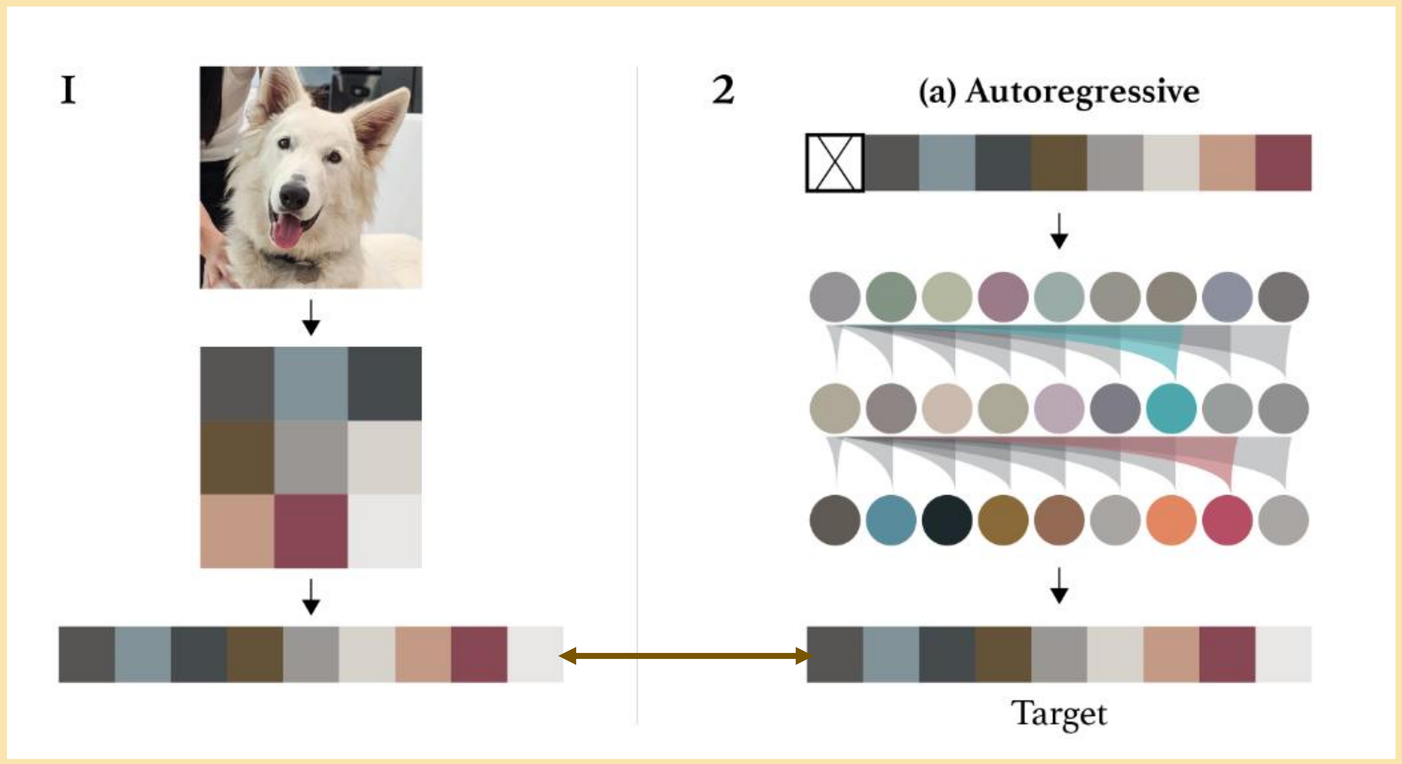
Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

Generation: An autoregressive view



Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

Generation: An autoregressive view

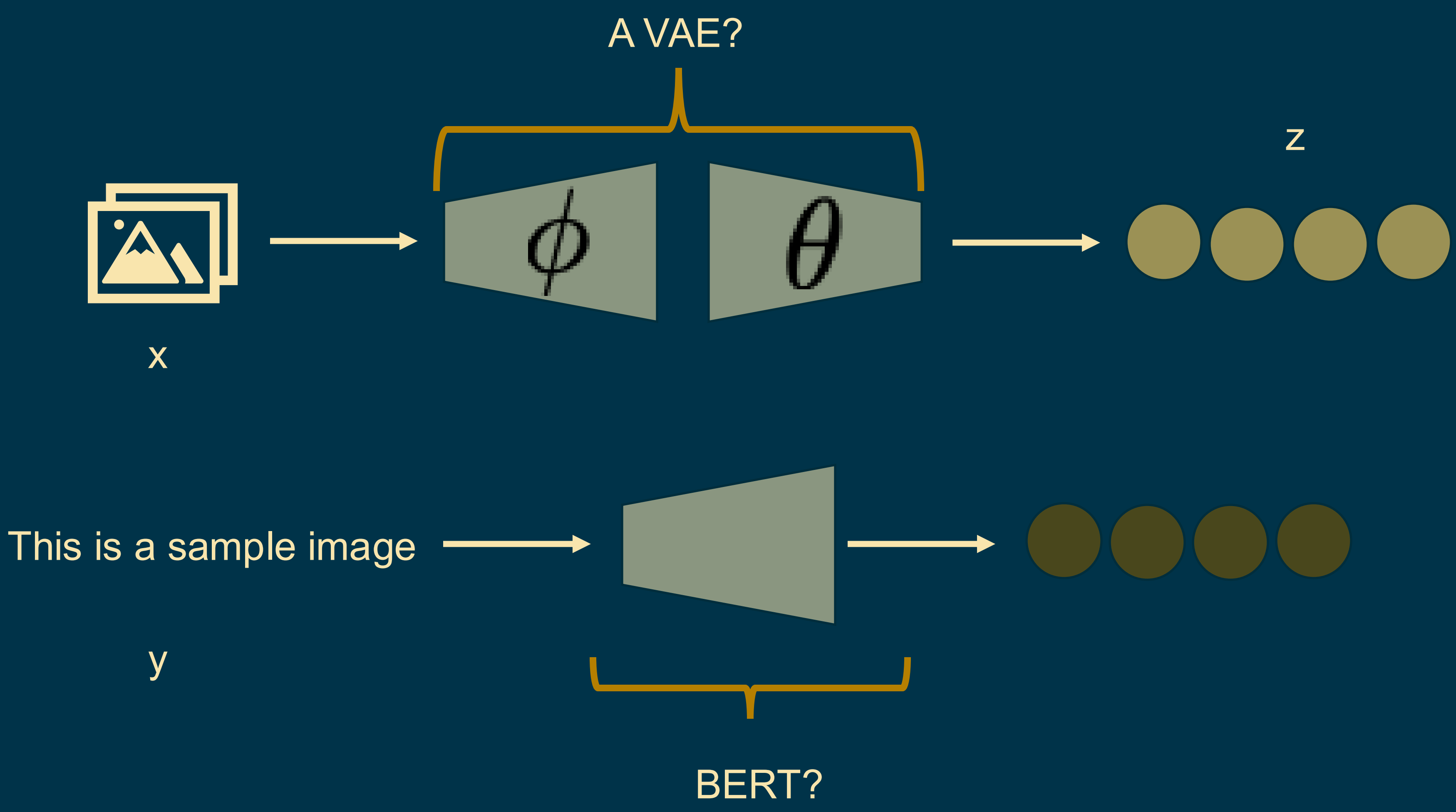


Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

Generation

But how do we include another modality?

Conditioning: A first attempt



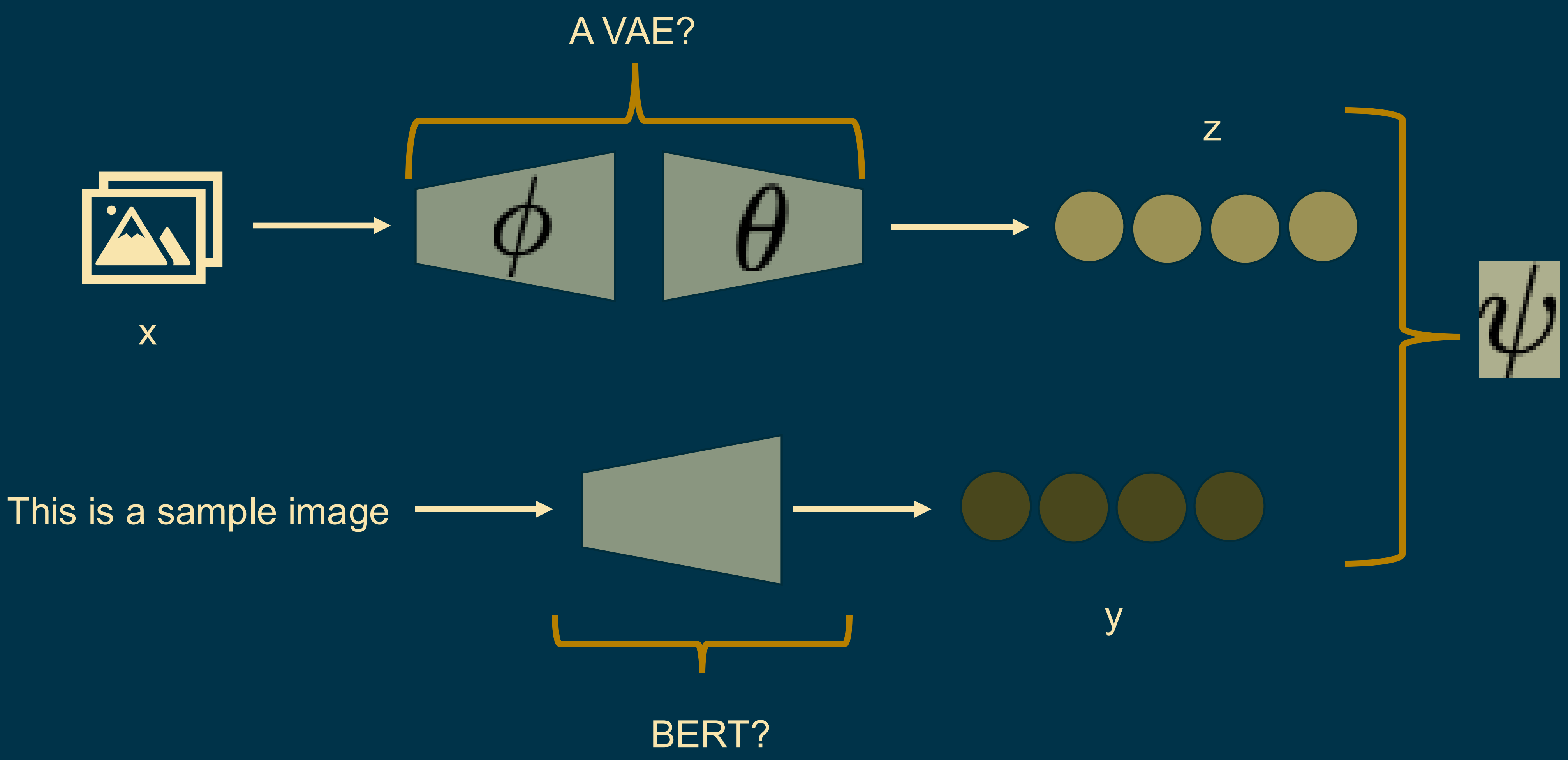
Conditioning: A first attempt

We have representations of the image and the text

Conditioning: A first attempt

We want to jointly model x, y, z

Conditioning: A first attempt



Conditioning: A first attempt

$$\ln p_{\theta,\psi}(x, y) \geq \mathbb{E}_{z\sim q_\phi(z|x)} \left(\underbrace{\ln p_\theta(x \mid y, z)}_{\text{VAE decoder}} - \beta D_{\mathrm{KL}}(q_\phi(y, z \mid x) \,\|\, p_\psi(y, z))\right)$$

Conditioning: A first attempt

$$\ln p_{\theta,\psi}(x, y) \geq \mathbb{E}_{z \sim q_\phi(z|x)} \left(\underbrace{\ln p_\theta(x \mid y, z)}_{\text{VAE decoder}} - \beta D_{\mathrm{KL}}(\underbrace{q_\phi(y, z \mid x)}_{\text{VAE encoder}} \,\|\, p_\psi(y, z)) \right)$$

Conditioning: A first attempt

$$\ln p_{\theta,\psi}(x, y) \geq \mathbb{E}_{z \sim q_\phi(z|x)} \left(\underbrace{\ln p_\theta(x \mid y, z)}_{\text{VAE decoder}} - \beta D_{\mathrm{KL}}\big(\underbrace{q_\phi(y, z \mid x)}_{\text{VAE encoder}} \,\|\, \underbrace{p_\psi(y, z)}_{\text{Transformer}}\big)\right)$$

Conditioning: A first attempt

$$\ln p_{\theta,\psi}(x,y) \geq \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}\left(\underbrace{\ln p_\theta(x \mid y,z)}_{\text{VAE decoder}} - \beta D_{\mathrm{KL}}(\underbrace{q_\phi(y,z \mid x)}_{\text{VAE encoder}} \,\|\, \underbrace{p_\psi(y,z)}_{\text{Transformer}})\right)}_{\text{ELBO}}$$

Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International conference on machine learning. Pmlr, 2021.

Conditioning: A first attempt

Two-stage training

First, train the VAE

Initial prior a uniform distribution over a K-codebook

=> Maximise ELBO wrt theta and phi

Conditioning: A first attempt

Two-stage training

First, train the VAE

Initial prior a uniform distribution over a K-codebook

=> Maximise ELBO wrt theta and phi

Next, train the transformer

=> Maximise ELBO wrt psi

Text and image concatenated as a single stream of data

Conditioning: A first attempt

Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International conference on machine learning. Pmlr, 2021.

Conditioning: A second attempt

Are there some issues with the autoregressive approach?

Conditioning: A second attempt

"….promising as our architecture uses a dense connectivity pattern which does not encode the 2D spatial structure of images yet is able to match and even outperform approaches which do…."

Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

Conditioning: A second attempt

"….promising as our architecture uses a dense connectivity pattern which does not encode the 2D spatial structure of images yet is able to match and even outperform approaches which do…."

Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

Conditioning: A second attempt

We need inductive bias

Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

Conditioning: A second attempt

Are there some issues with the autoregressive approach?

Conditioning: A second attempt

High-dimensional modelling in pixel-space is extremely inefficient

Conditioning: A second attempt

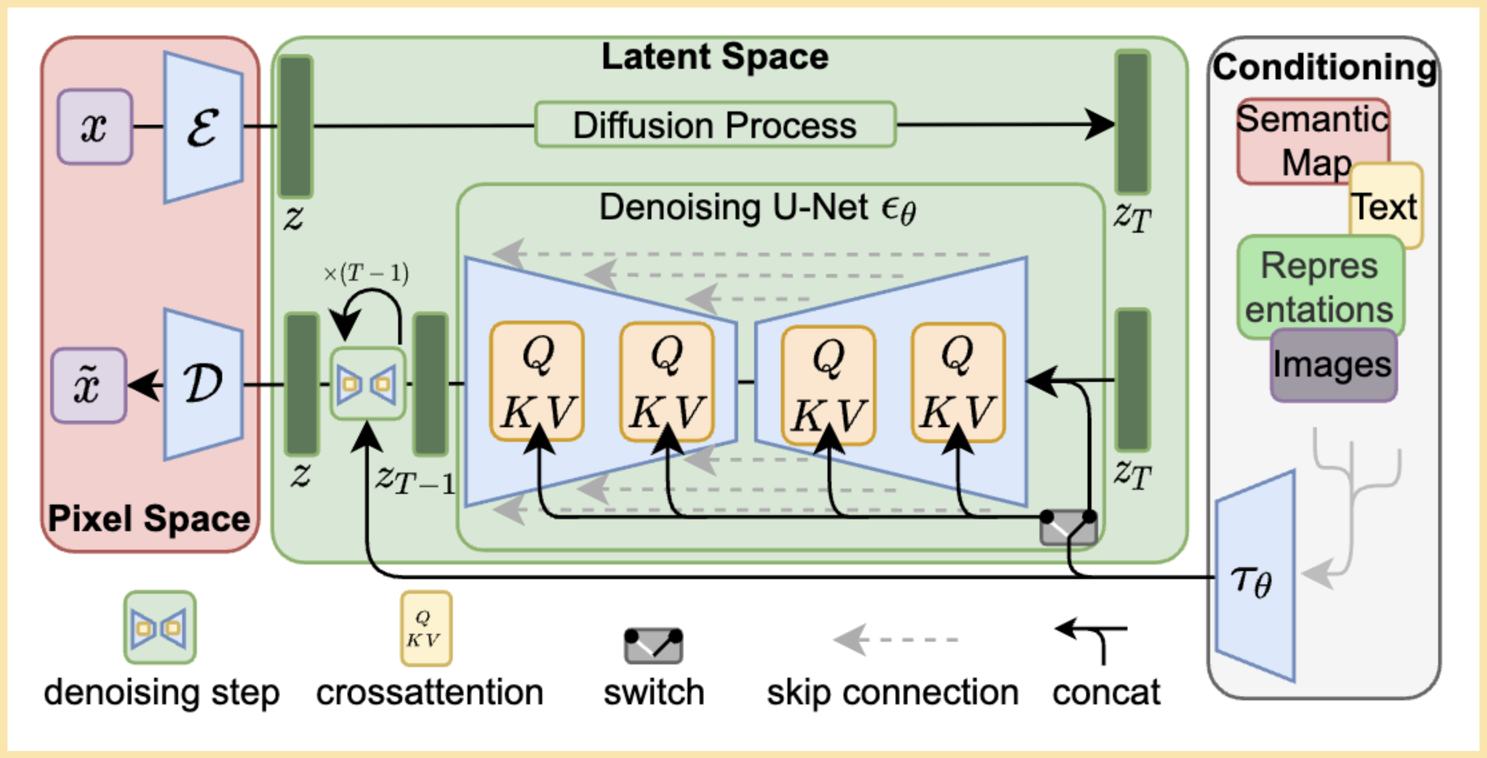
U-Net and 2D convolutions to preserve spatial structure

Compressed latent space with a VAE backbone

Cross-attention between text and image embeddings

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.

Conditioning: A second attempt



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.

Conditioning: A second attempt

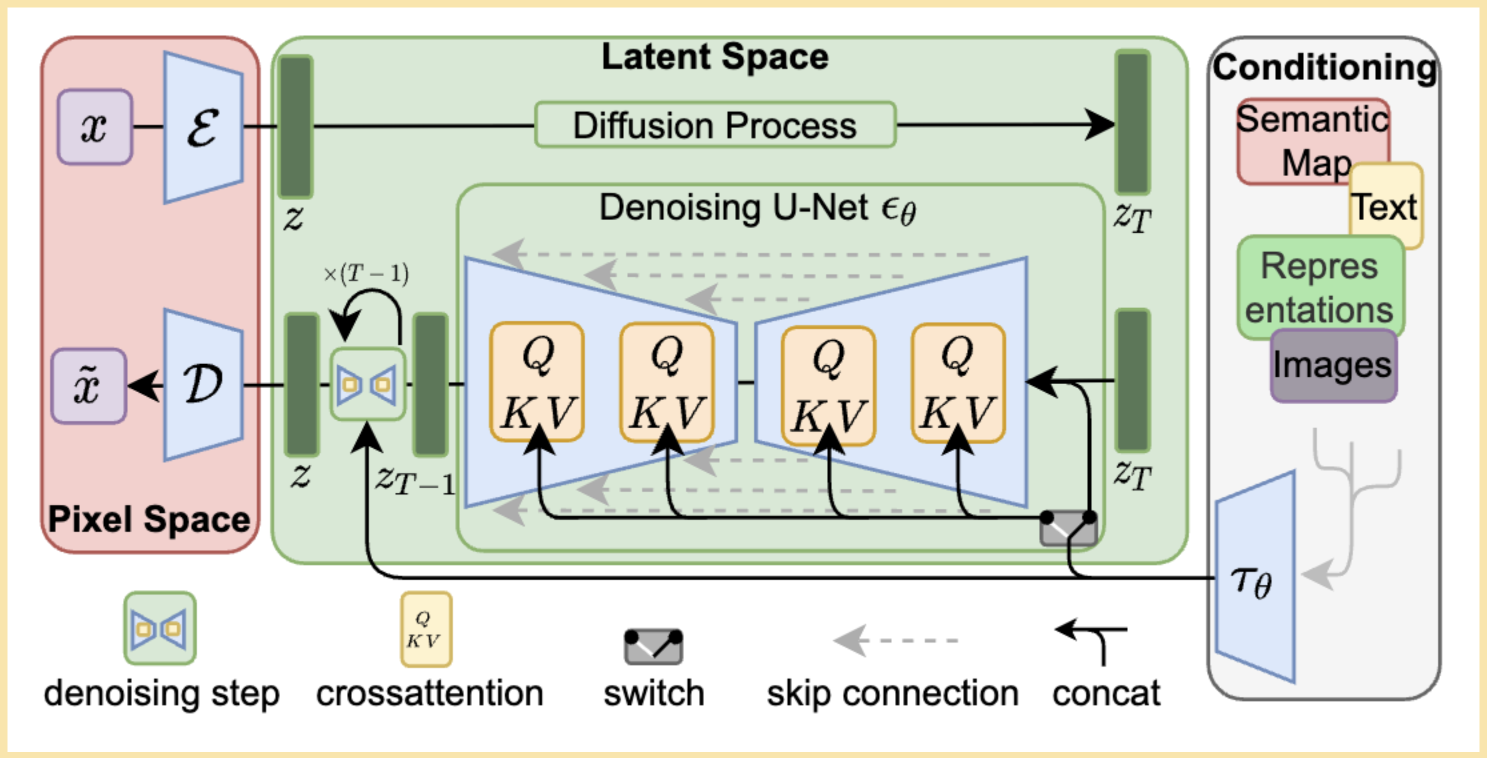
DDIM

$$\mathbb{E}_{\mathcal{E}(x),\, y,\, \epsilon \sim \mathcal{N}(0,1),\, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

Denoising U-Net Sparse transformer

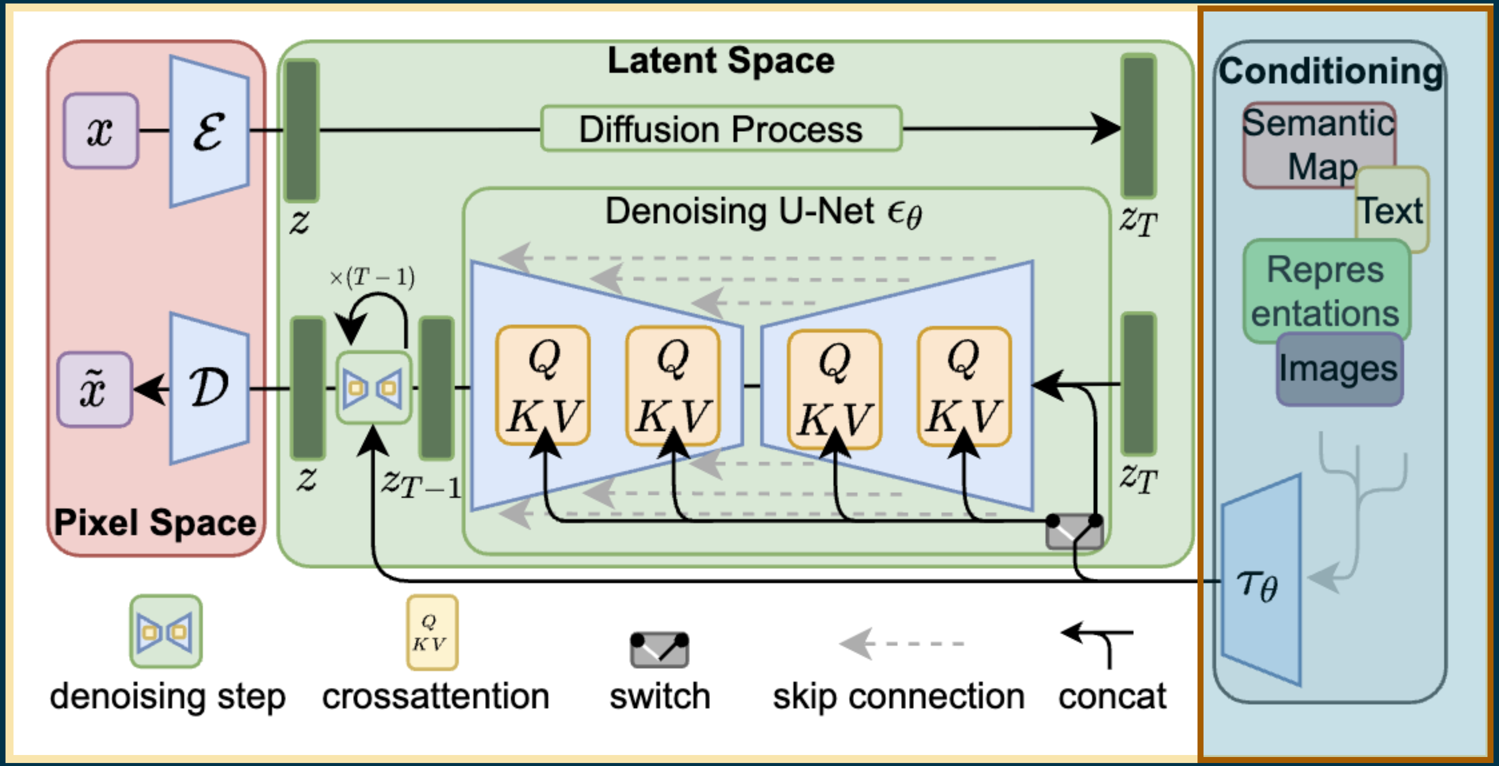
Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Conditioning Zoo



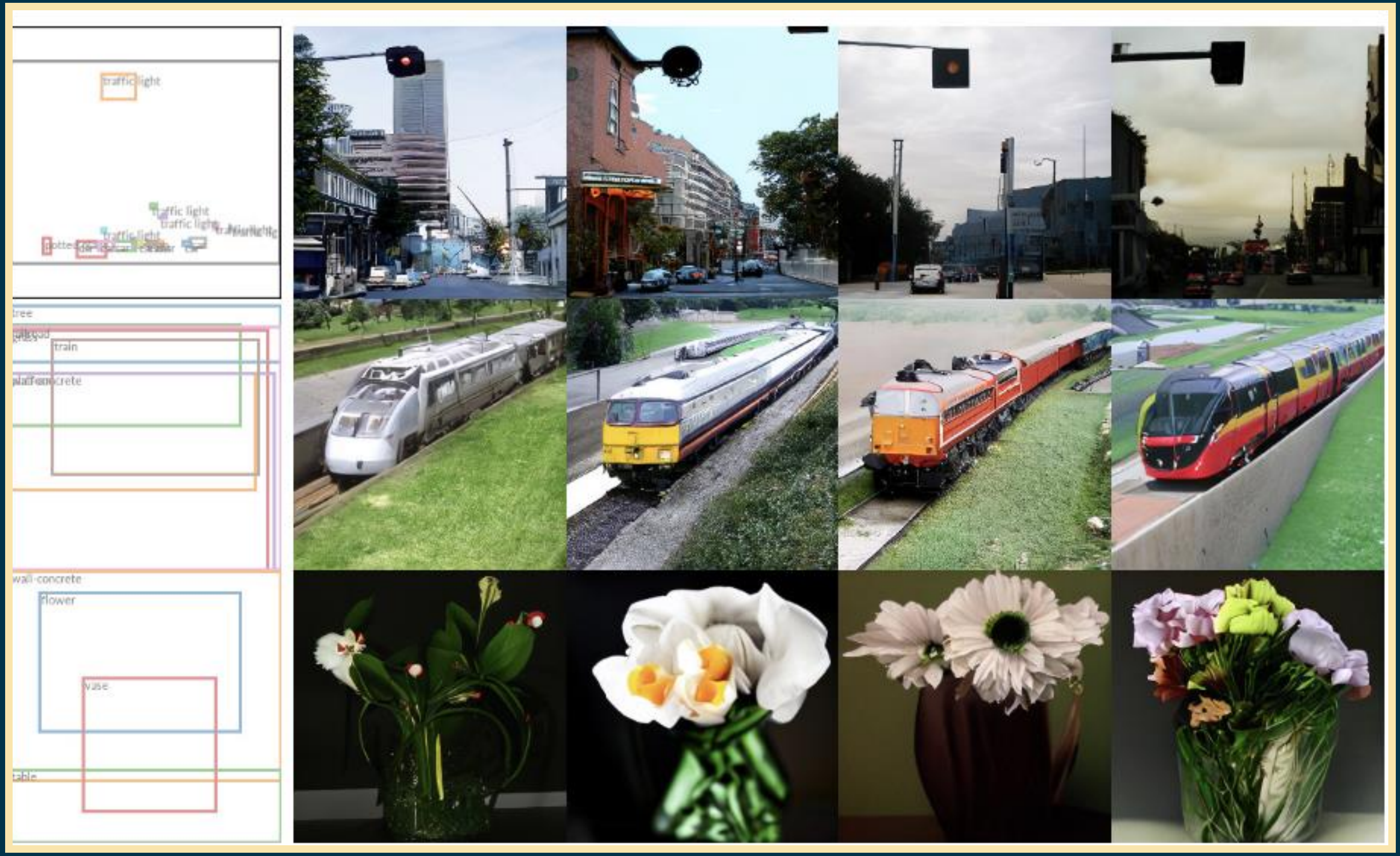
Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Conditioning Zoo



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.

Beyond Text

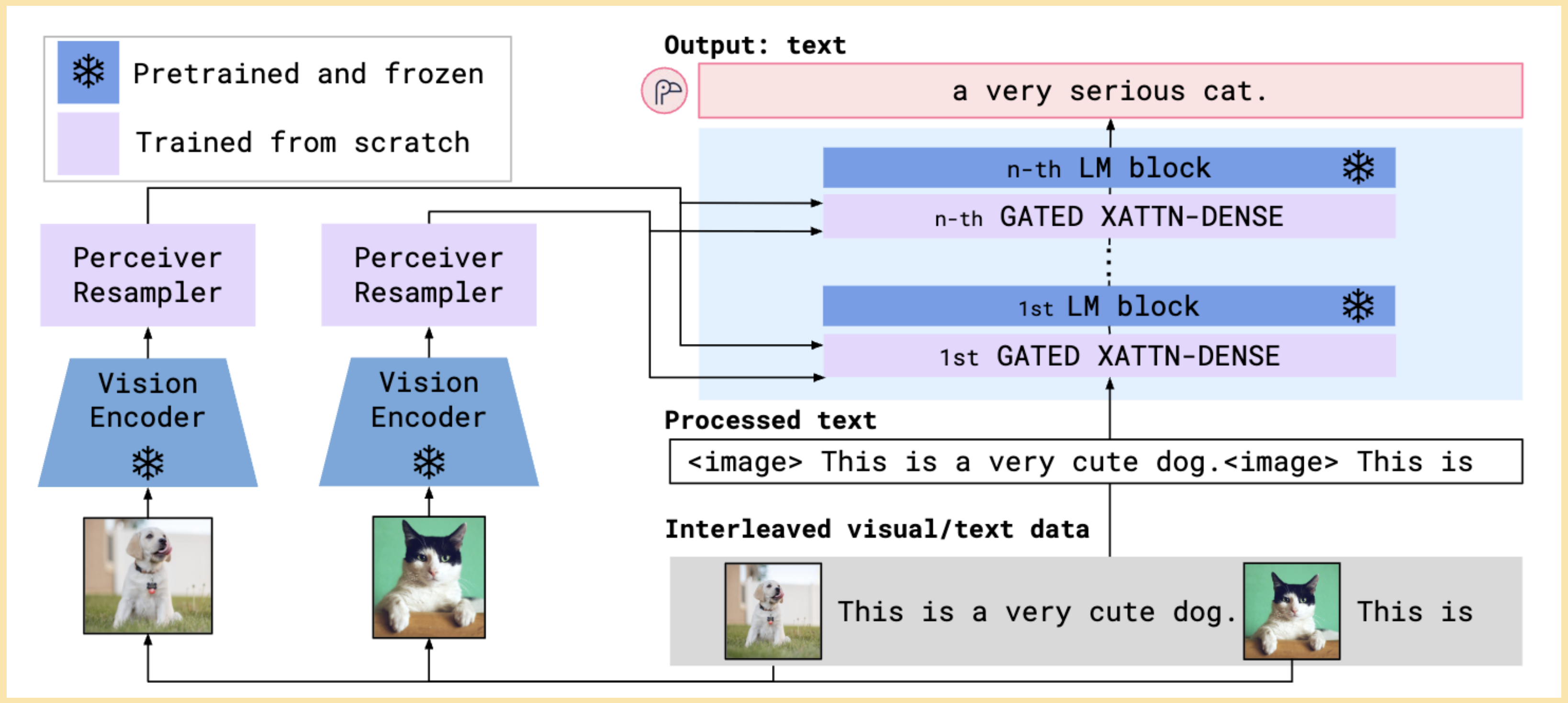


Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Beyond Text

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.

Generative vs Discriminative: A blurry line



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

In this talk

Conditioning

Control

Edits

Cool Stuff

Control: A form of fine-tuning

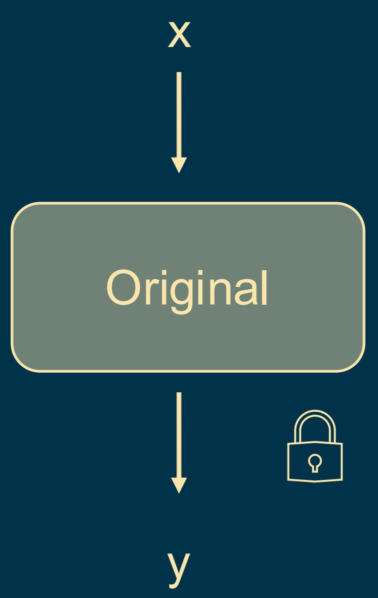
Given a pretrained text-to-image model

Can we generate new, spatially localized, task specific images?

Control: A form of fine-tuning

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Control: A residual on a clone

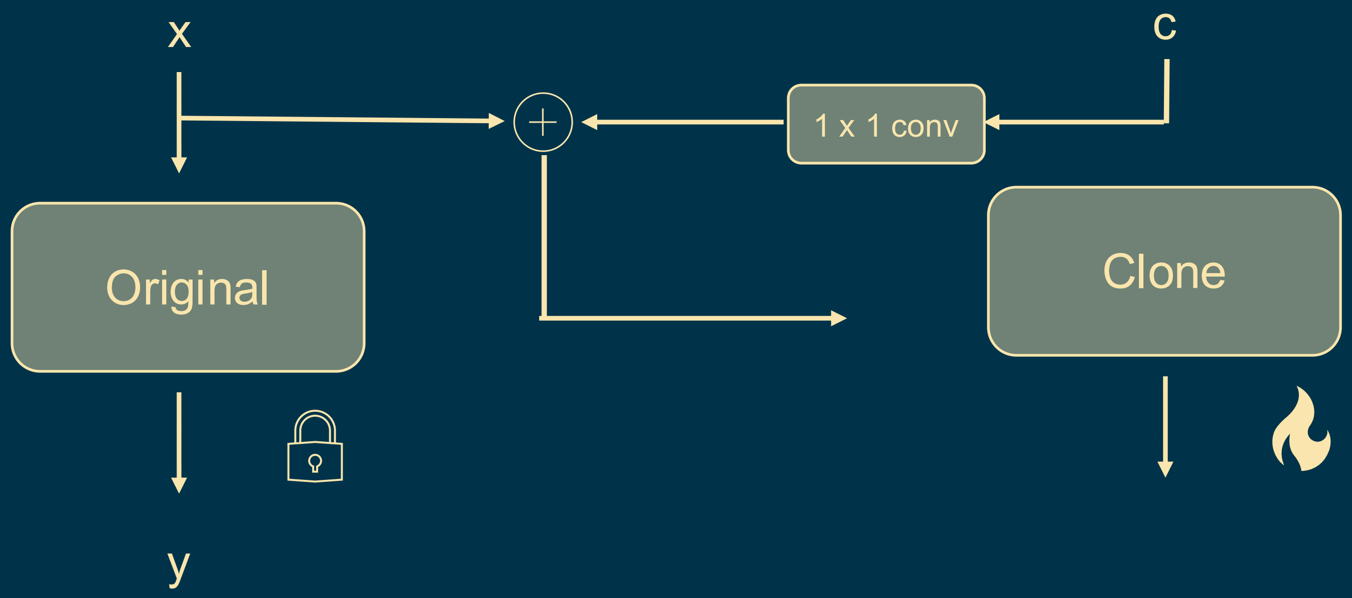


Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Control: A residual on a clone

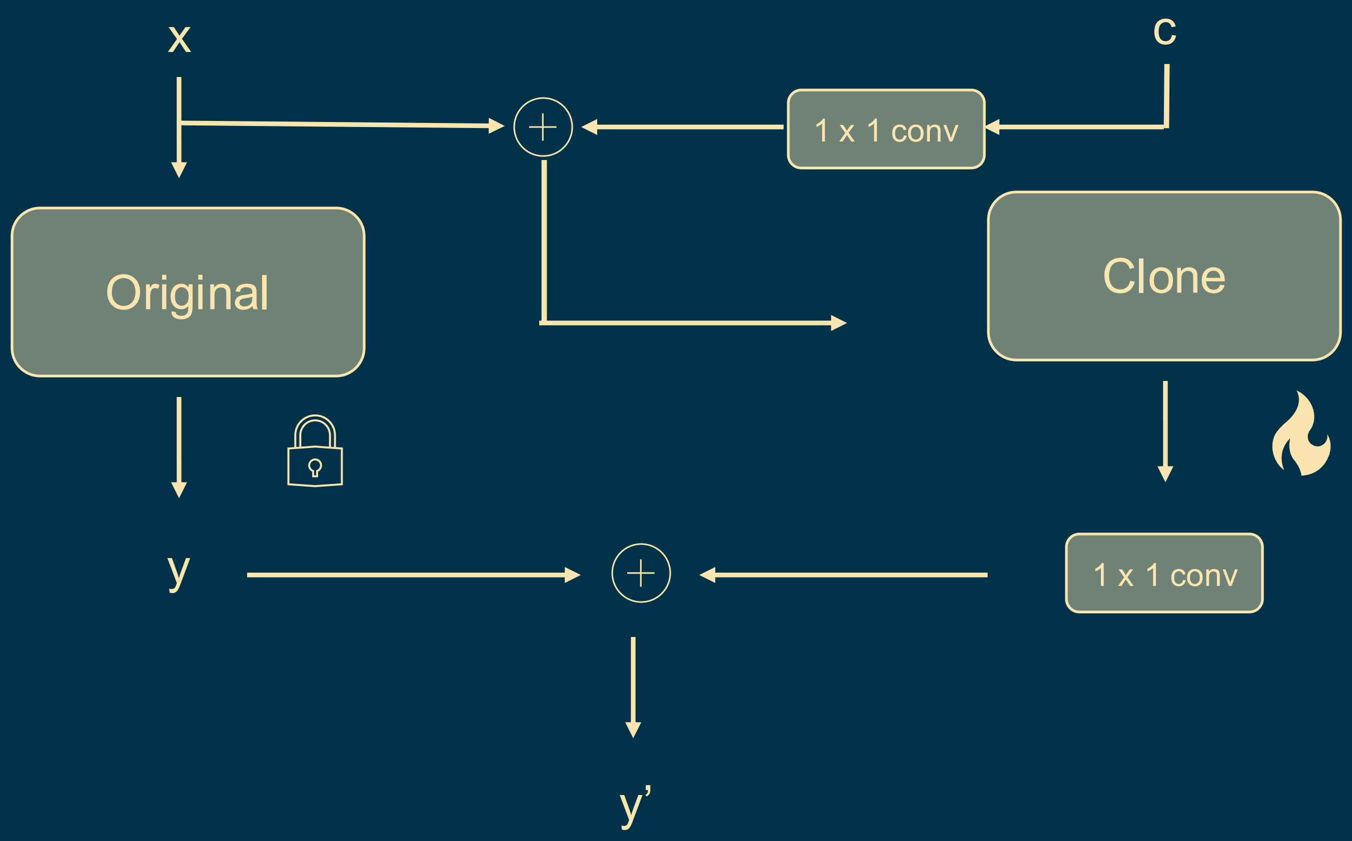
Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Control: A residual on a clone



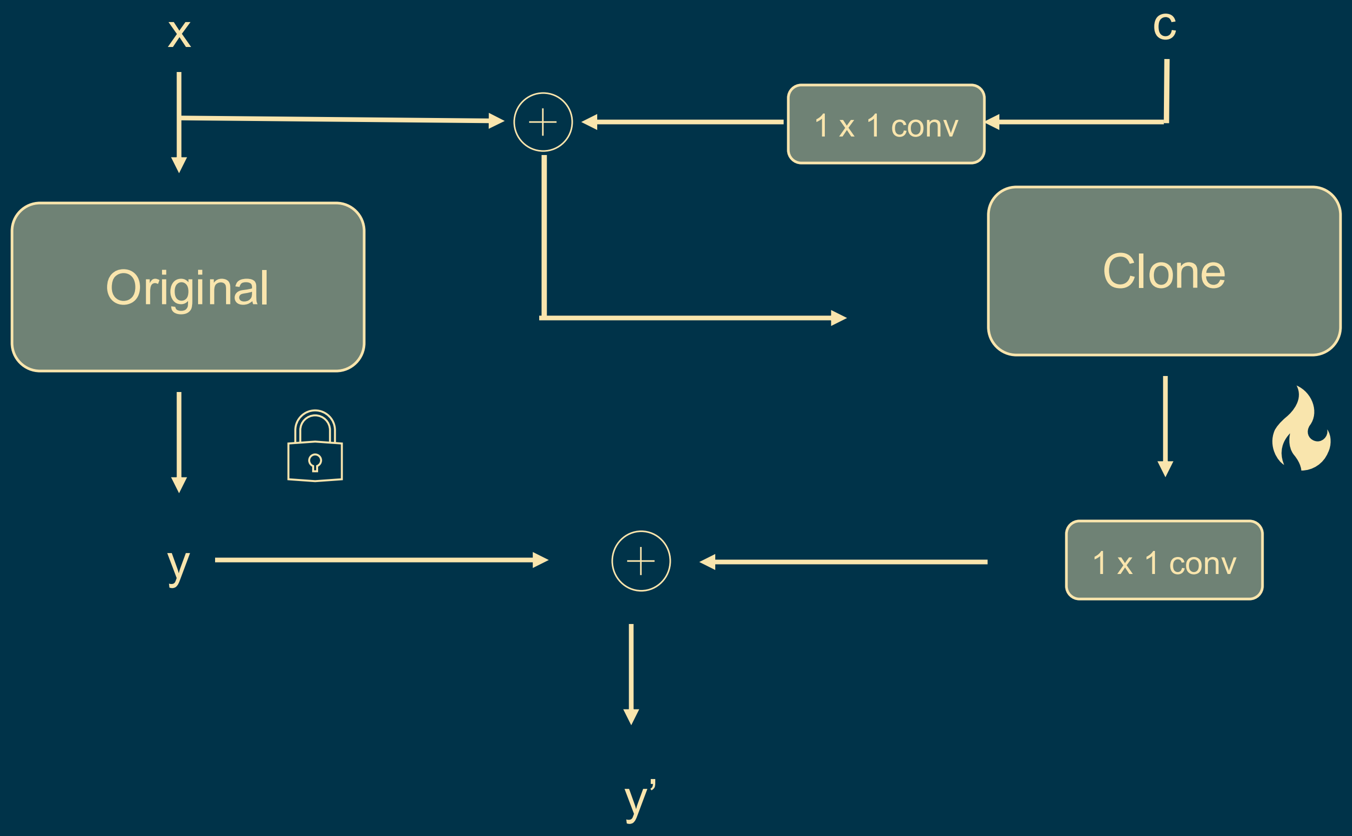
Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Control: A residual on a clone



Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Control: A residual on a clone



Why a *zero convolution?*

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

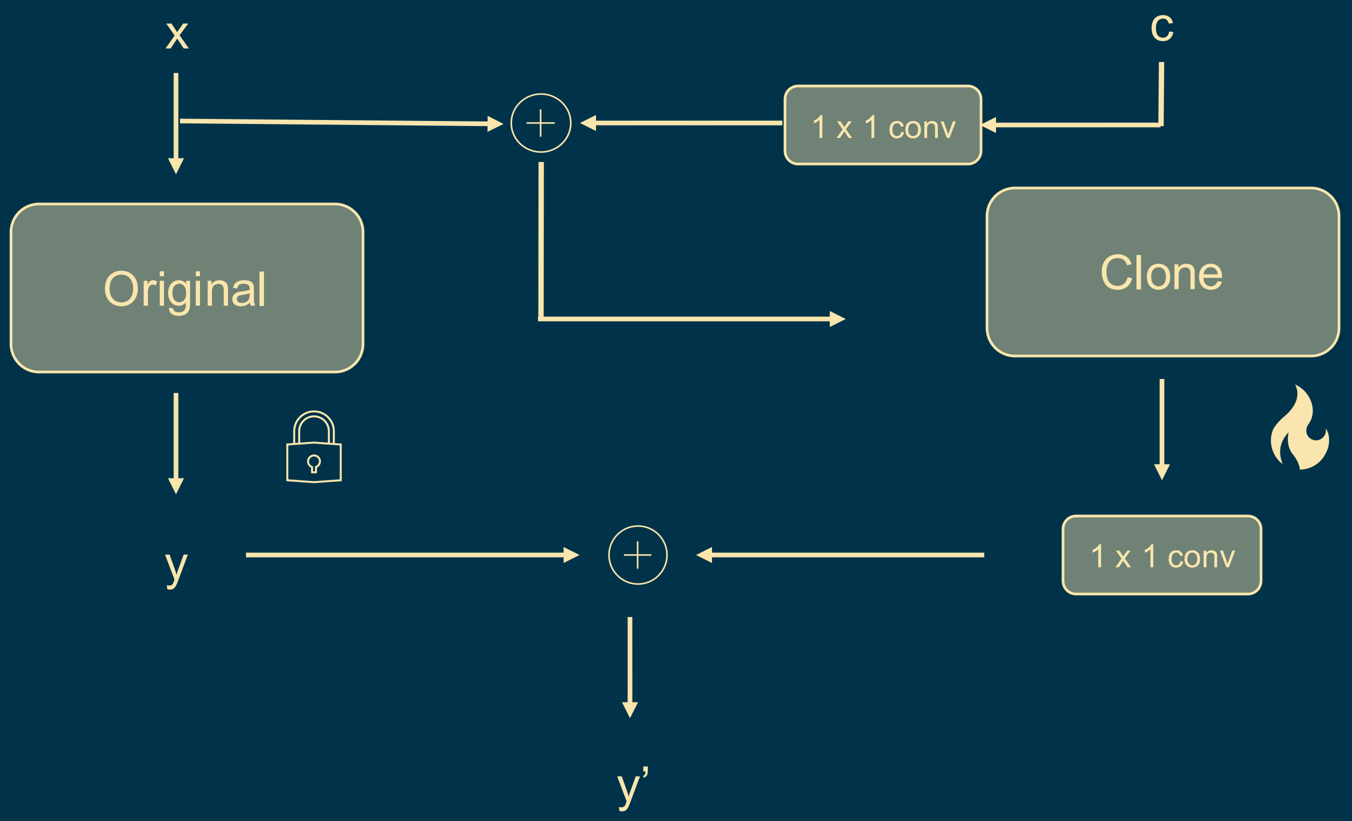
The Zero Convolution

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

What happens when zero?

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Control: A residual on a clone

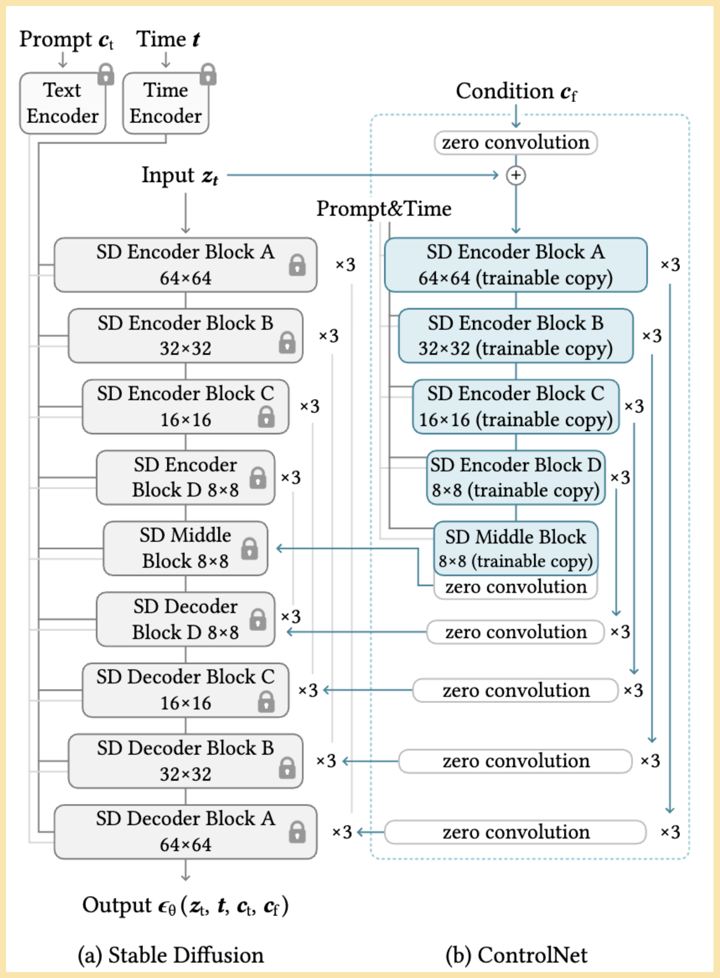


Why a *zero convolution?*

In the first forward pass, keep things unchanged

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

ControlNet



Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

ControlNet

$$\mathbb{E}_{\mathcal{E}(x),\, y,\, \epsilon\sim\mathcal{N}(0,1),\, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

Before

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

ControlNet

$$\mathbb{E}_{\mathcal{E}(x),\, y,\, \epsilon\sim\mathcal{N}(0,1),\, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

Before

$$\mathbb{E}_{z_0,\, t,\, c_t,\, c_f,\, \epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, \mathbf{t}, \mathbf{c}_t, c_f)\|_2^2\right]$$

After

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

ControlNet

Text prompt

$$\mathbb{E}_{\mathcal{E}(x),\, y,\, \epsilon\sim\mathcal{N}(0,1),\, t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\Big]$$

Before

$$\mathbb{E}_{z_0,\, t,\, c_t,\, c_f,\, \epsilon\sim\mathcal{N}(0,1)}\Big[\|\epsilon - \epsilon_\theta(z_t, \mathbf{t}, \mathbf{c}_t, c_f)\|_2^2\Big]$$

After

Task vector

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

ControlNet: Text not a necessity!

$$\mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(z_t, \mathbf{t}, \mathbf{c}_t, c_f) \|_2^2 \right]$$

Zero!

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

In this talk

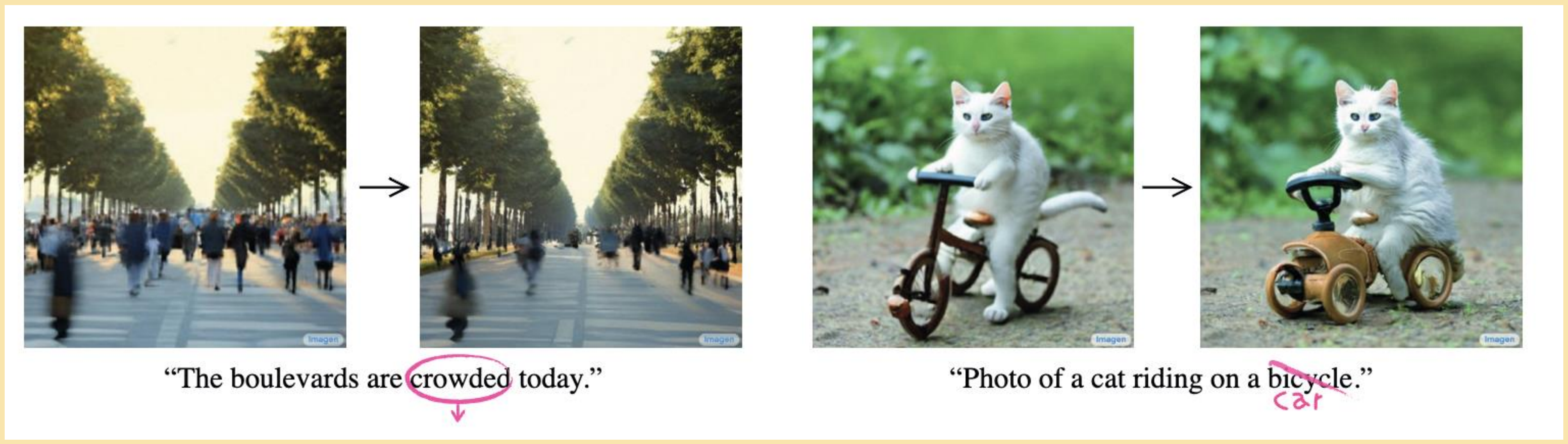
Conditioning

Control

Edits

Cool Stuff

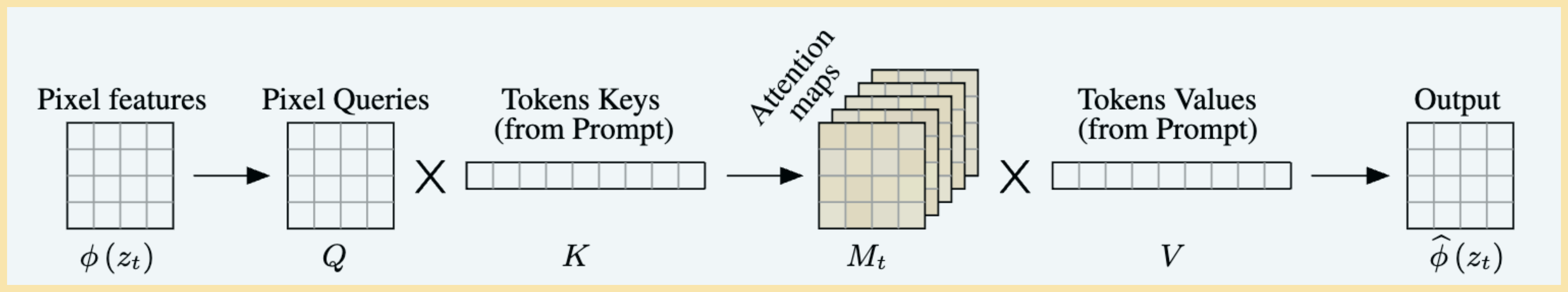
Text driven image editing



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

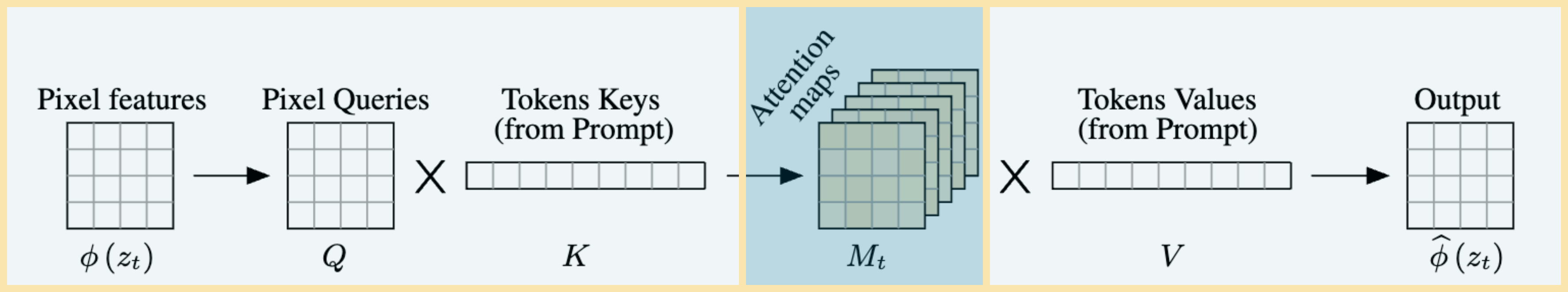
The cross-attention layers hold the information we need



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

A photo of a cat on a bicycle ⟶

Prompt P

Attention map M

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

A photo of a cat on a bicycle → A photo of a cat on a car

Prompt P → Prompt P*

Attention map M → Attention map M*

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

Preserve original layout and geometry

A photo of a cat on a bicycle → A photo of a cat on a car

Prompt P

Attention map M

Prompt P*

Attention map M*

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

A photo of a cat on a bicycle

Prompt P

Attention map M

→

A photo of a cat on a car

Prompt P*

Attention map M*

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

A photo of a cat on a bicycle

Prompt P ✗

Attention map M ✓

→

A photo of a cat on a car

Prompt P* ✓

Attention map M* ✗

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

A photo of a cat on a bicycle

Prompt P ✗

Attention map M ✓

A photo of a cat on a car

Prompt P* ✓

Attention map M* ✗

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing – Phrases

The cross-attention layers hold the information we need

A photo of a car on the side of the street → A photo of a sports car on the side of the street

Prompt P

Attention map M

Prompt P*

Attention map M*

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing – Phrases

The cross-attention layers hold the information we need

A photo of a car on the side of the street ⟶ A photo of a sports car on the side of the street

Prompt P

Prompt P*

Add only the modified maps

Attention map M ⟷ Attention map M*

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing – Phrases

The cross-attention layers hold the information we need

A photo of a car on the side of the street → A photo of a sports car on the side of the street

Prompt P

Attention map M

Add only the modified maps

Everything else unchanged

Prompt P*

Attention map M*

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

The cross-attention layers hold the information we need

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Text driven image editing

How about instructions?

Text driven image editing

Instead of:

A photo of the Eiffel Tower → A photo of the Eiffer tower with fireworks in the sky

Text driven image editing

Instead of:

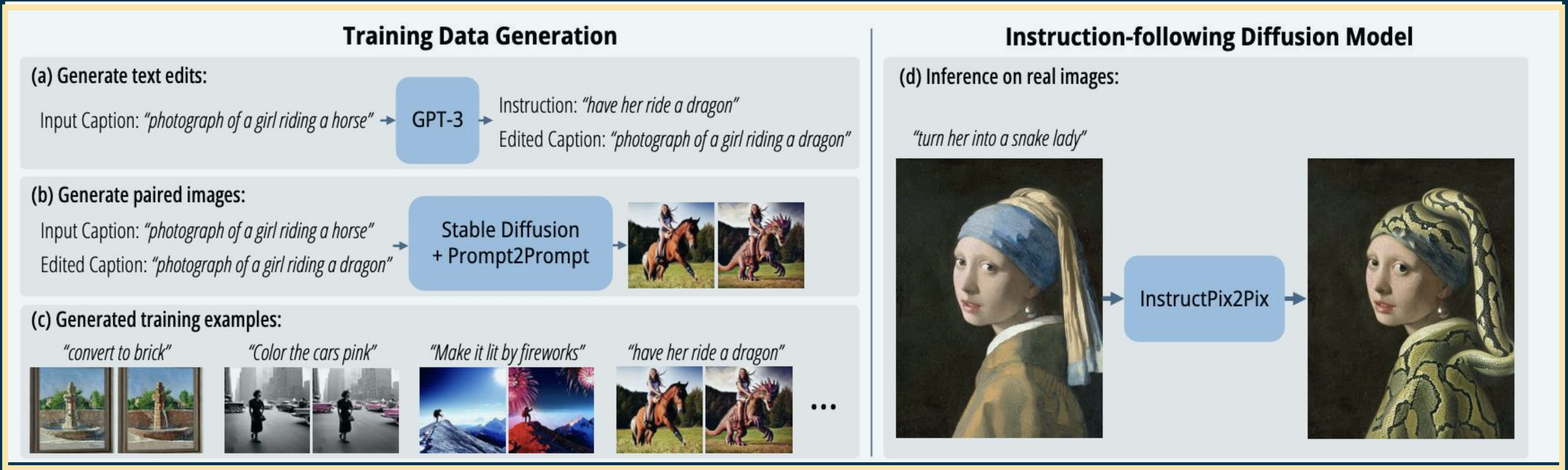
A photo of the Eiffel Tower ⟶ A photo of the Eiffer tower with fireworks in the sky

How about:

Add fireworks to the sky

Text driven image editing

Instruction following as supervised learning



Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

In this talk

Conditioning

Control

Edits

Cool Stuff

Cool Stuff

Think about all of this as representation learning

We can use representations to do *anything*

Cool Stuff

Think about all of this as representation learning

So why not couple generative and discriminative models?

Cool Stuff

Recall that CLIP can classify images into labels

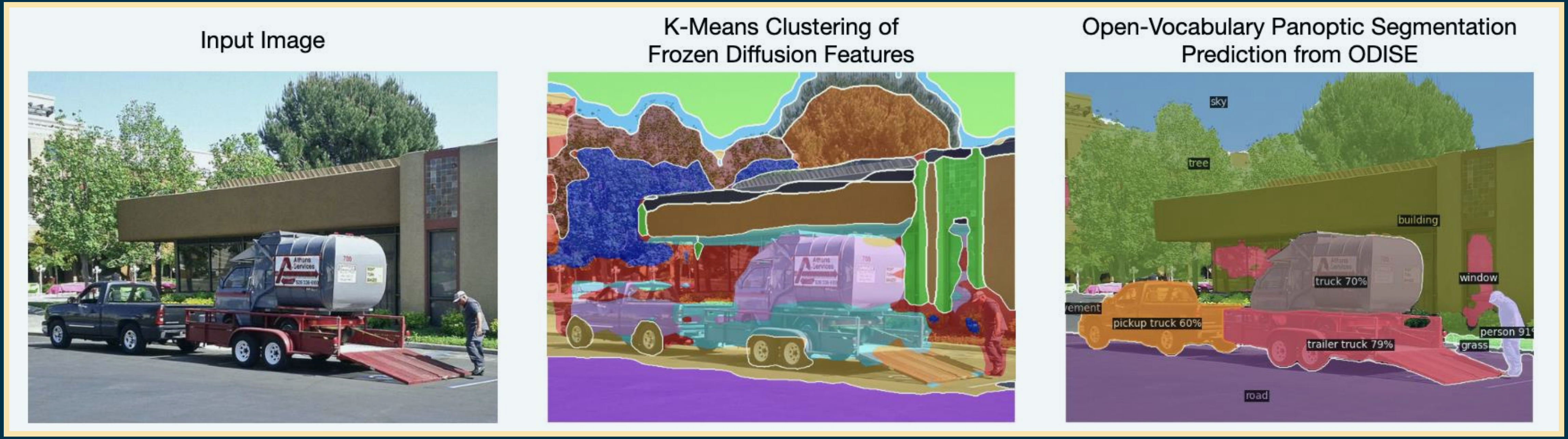
We now know that T2I models learn rich representations of world concepts

Cool Stuff

Recall that CLIP can classify images into labels

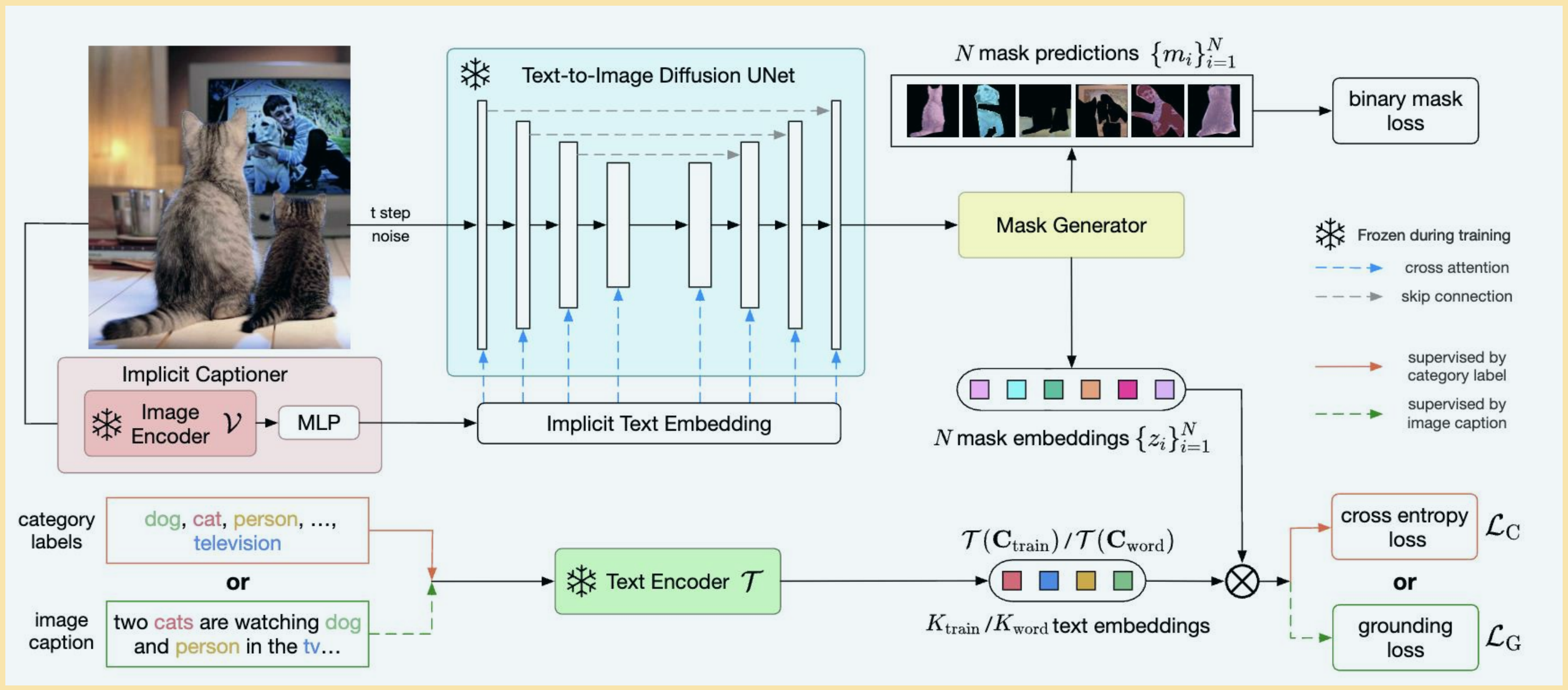
We now know that T2I models learn rich representations of world concepts

Open World Segmentation



Xu, Jiarui, et al. "Open-vocabulary panoptic segmentation with text-to-image diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Open World Segmentation

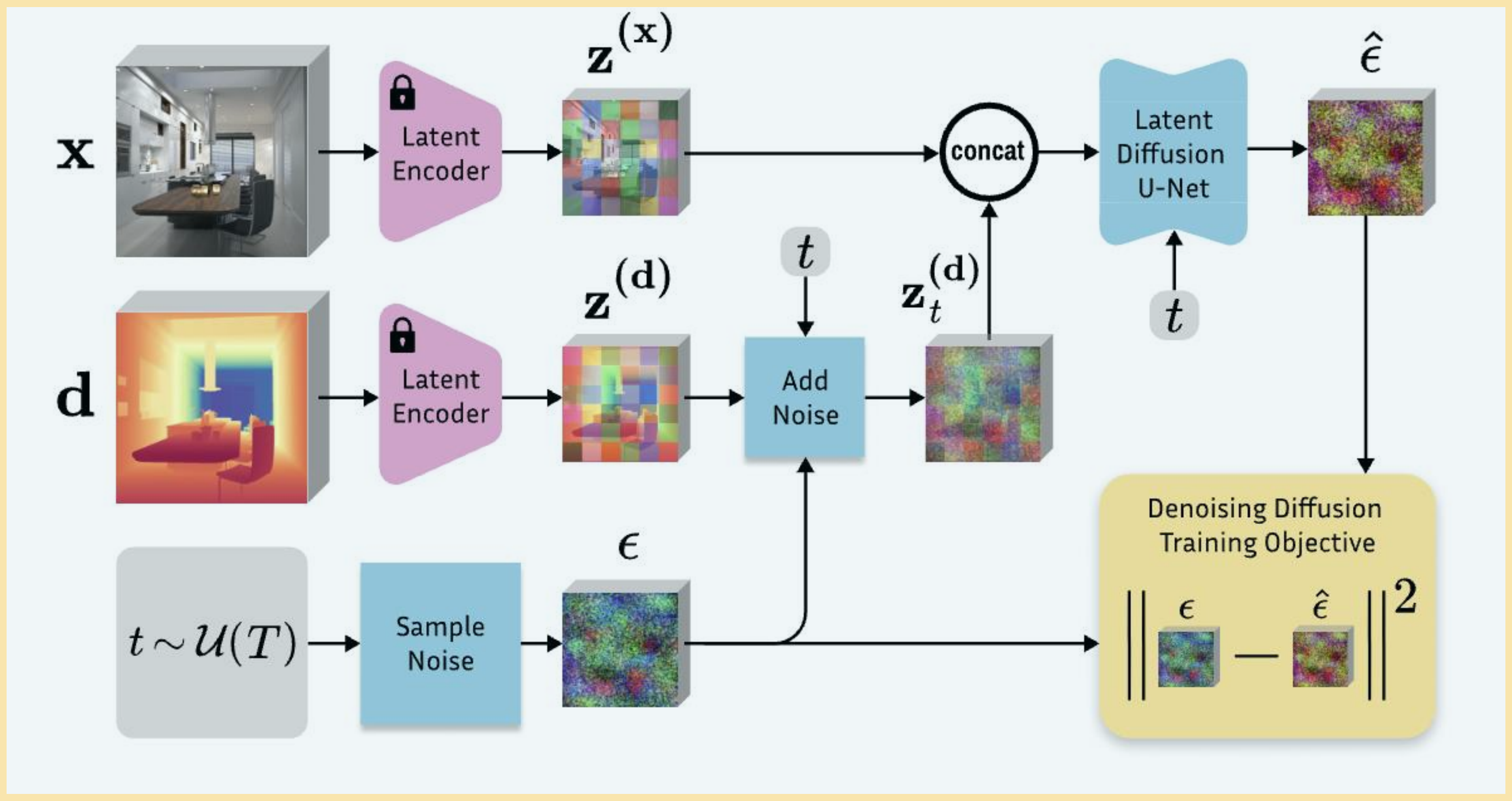


Xu, Jiarui, et al. "Open-vocabulary panoptic segmentation with text-to-image diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Depth Estimation

Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Estimation



Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Summary

Conditioning

Control

Edits

Cool Stuff

Incorporate a modality

Attention is your friend

Summary

Conditioning

Incorporate a modality

Attention is your friend

Control

Fine-tune your generation

Text is not a necessity

Edits

Cool Stuff

Summary

Conditioning
- Incorporate a modality
- Attention is your friend

Control
- Fine-tune your generation
- Text is not a necessity

Edits
- Follow prompts
- Follow instructions

Cool Stuff

Summary

Conditioning	Control	Edits	Cool Stuff
Incorporate a modality	Fine-tune your generation	Follow prompts	Do anything with frozen representations
Attention is your friend	Text is not a necessity	Follow instructions	