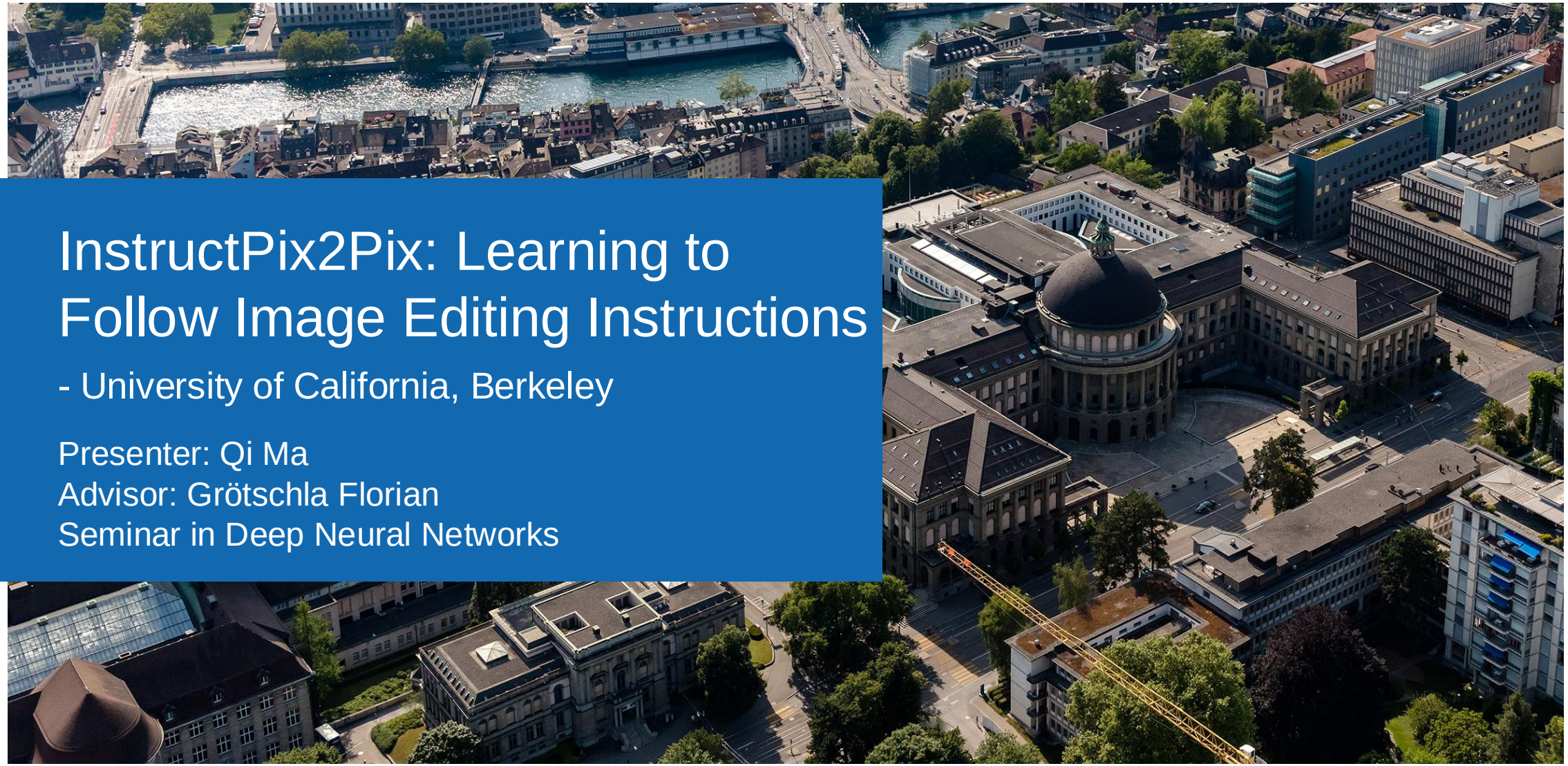
ETH zürich

CVL Computer Vision Lab



InstructPix2Pix: Learning to Follow Image Editing Instructions

- University of California, Berkeley

Presenter: Qi Ma
Advisor: Grötschla Florian
Seminar in Deep Neural Networks

Motivation :

Motivation:

- What kind of text-based image editing functionality do you expect to use?

Motivation:

- What kind of text-based image editing functionality do you expect to use?

Motivation:

- What kind of text-based image editing functionality do you expect to use?

Remove the people

Motivation:

- What kind of text-based image editing functionality do you expect to use?

Motivation:

- What kind of text-based image editing functionality do you expect to use?

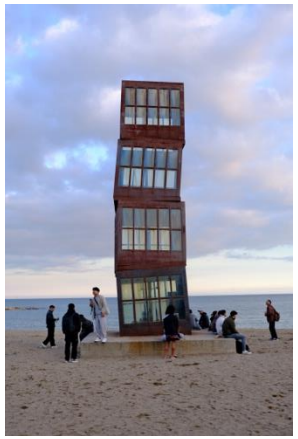


Image input:

Remove the people

Text input:

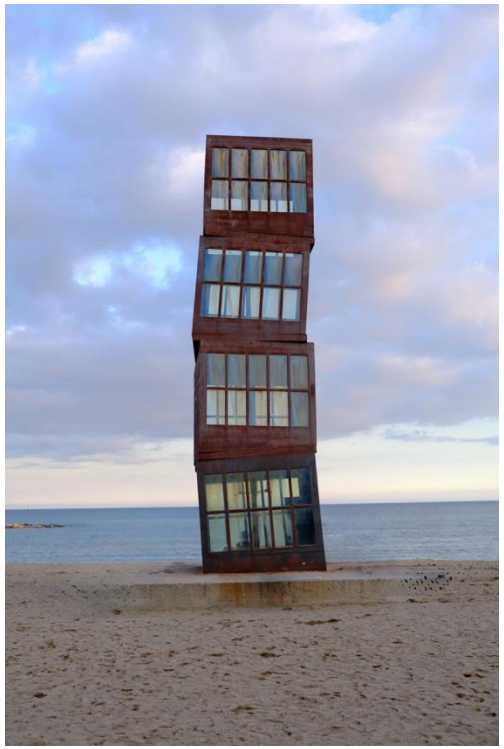


Image output:

Motivation:

- Sound easy? Why not try stable diffusion[1]?

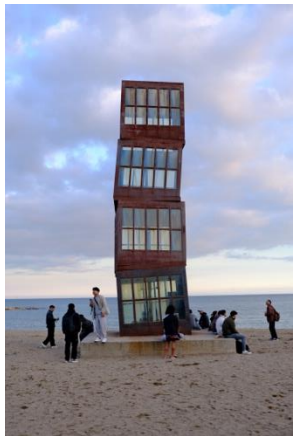
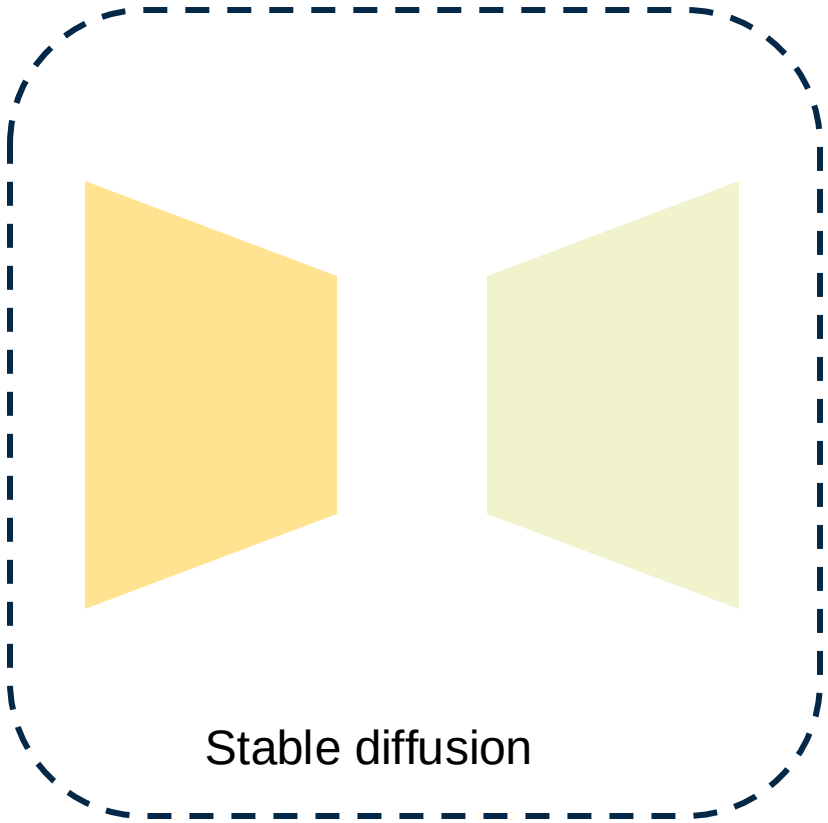


Image input:

Remove the people

Text input:



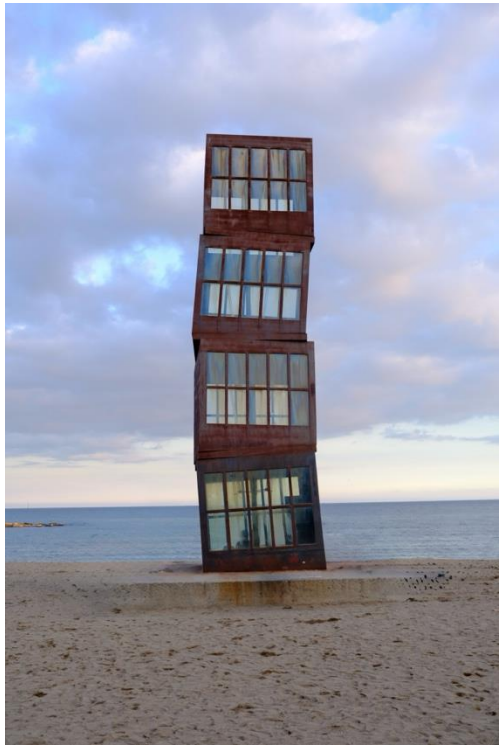
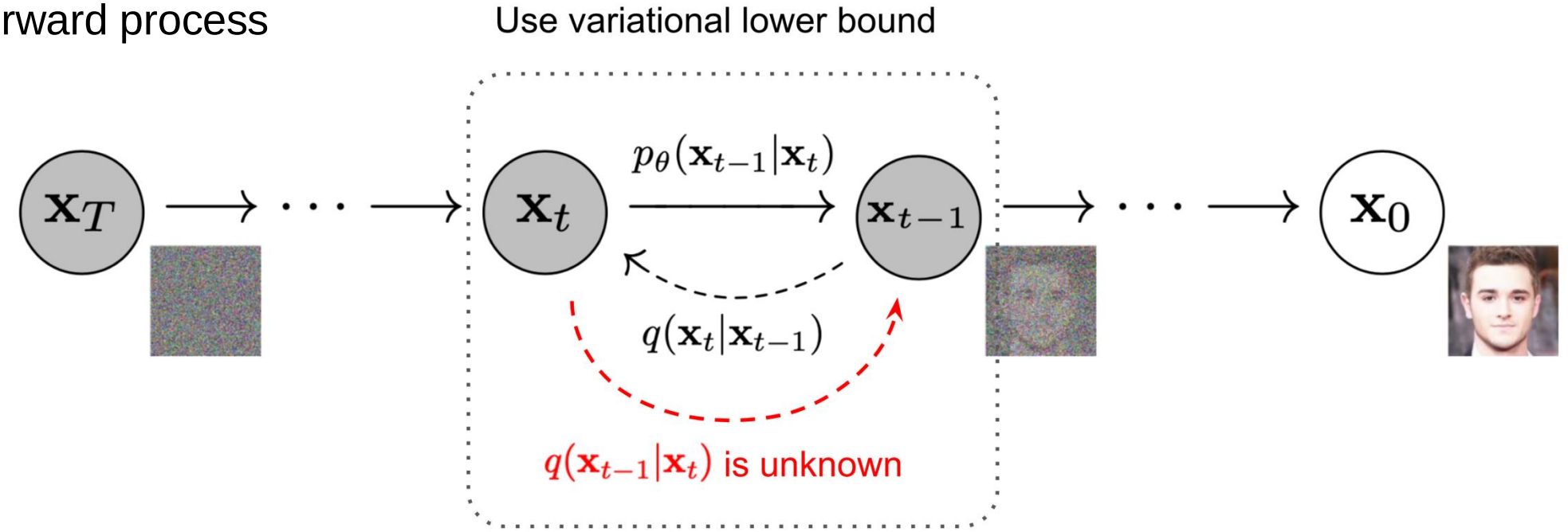


Image output:

[1] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Motivation:

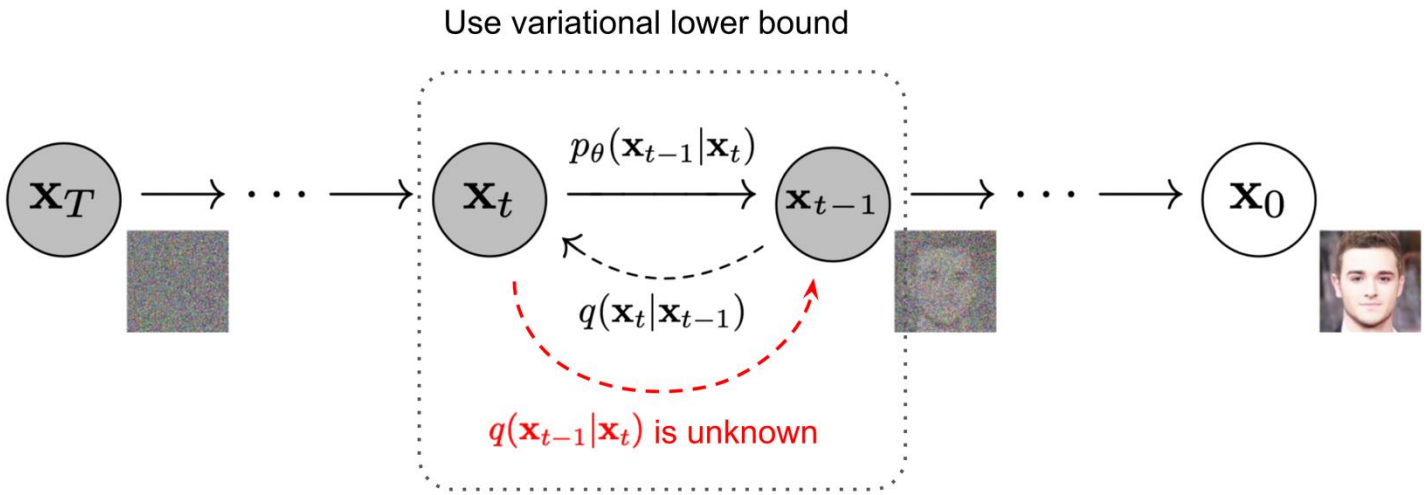
- What is diffusion model?
  - Forward process

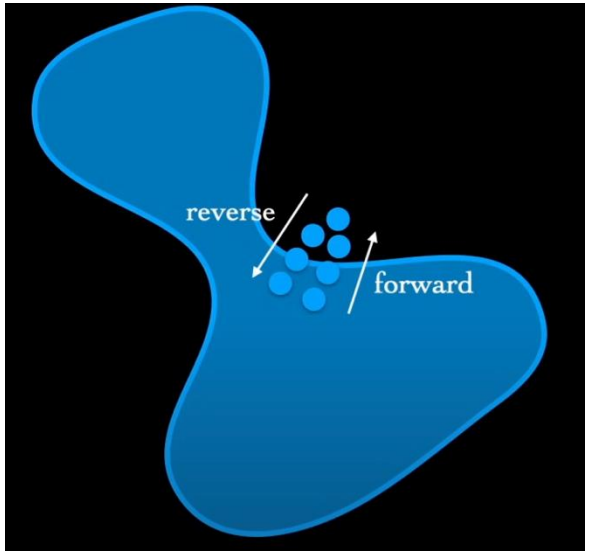


[2] https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Motivation:

- What is diffusion model?
  - Reverse process

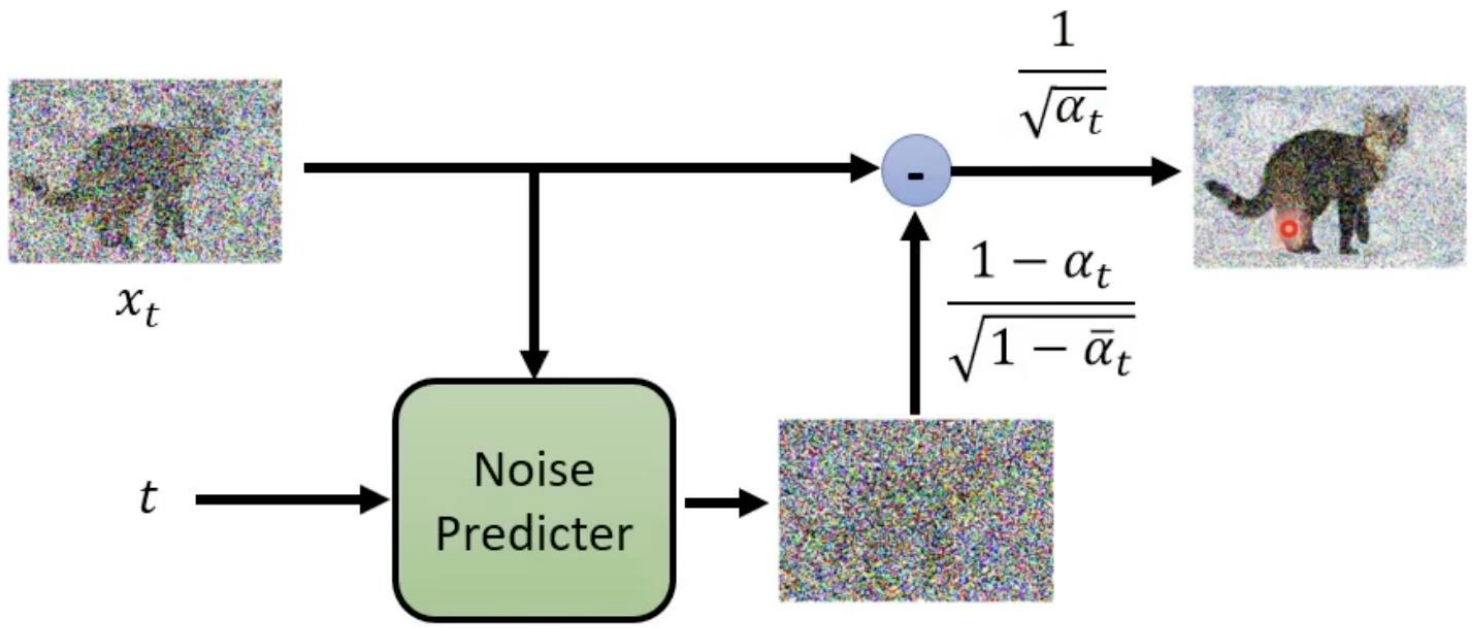


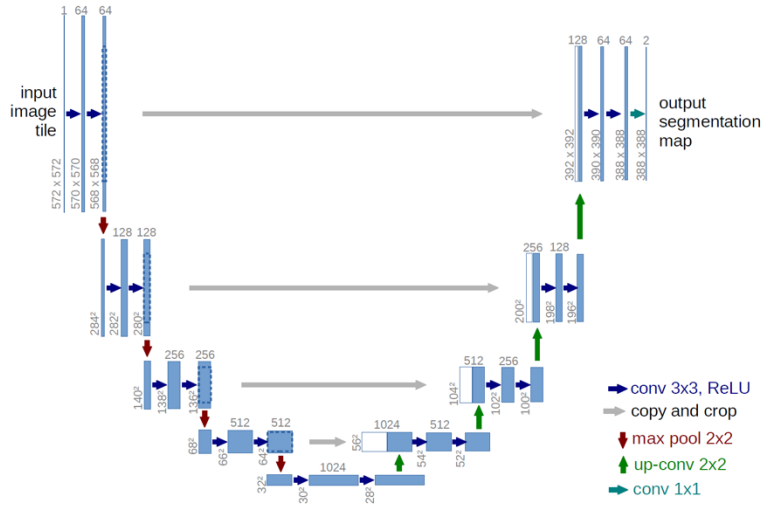


[3] https://www.youtube.com/watch?v=fbLgFrlTnGU

Motivation:

- What is diffusion model?



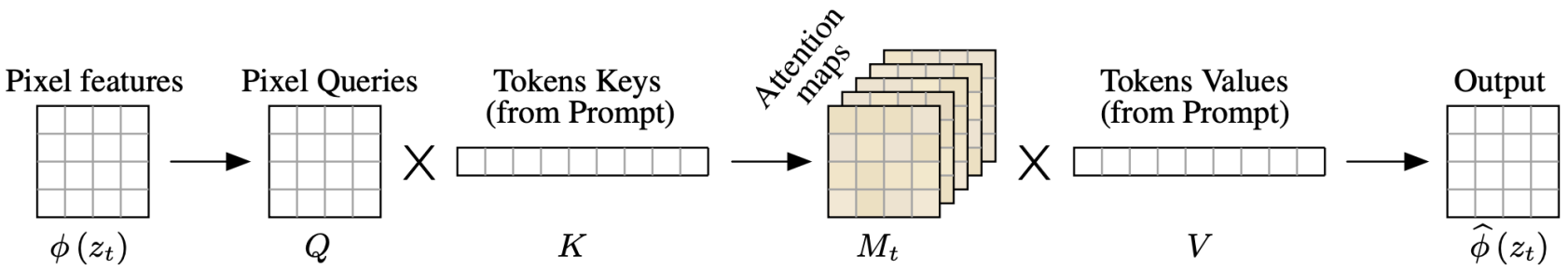


[3] https://www.youtube.com/watch?v=fbLgFrlTnGU

Motivation:

- How to condition on text prompt?
  - Cross attention operation in noise predicter

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V,$$



Q: (Token num, Feature_dim)

K: (Token num', Feature_dim)

M: (Token num x Token num')

[3] https://www.youtube.com/watch?v=fbLgFrlTnGU

Motivation:

- Sound easy? Why not try stable diffusion?

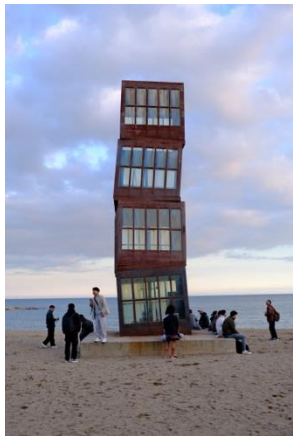
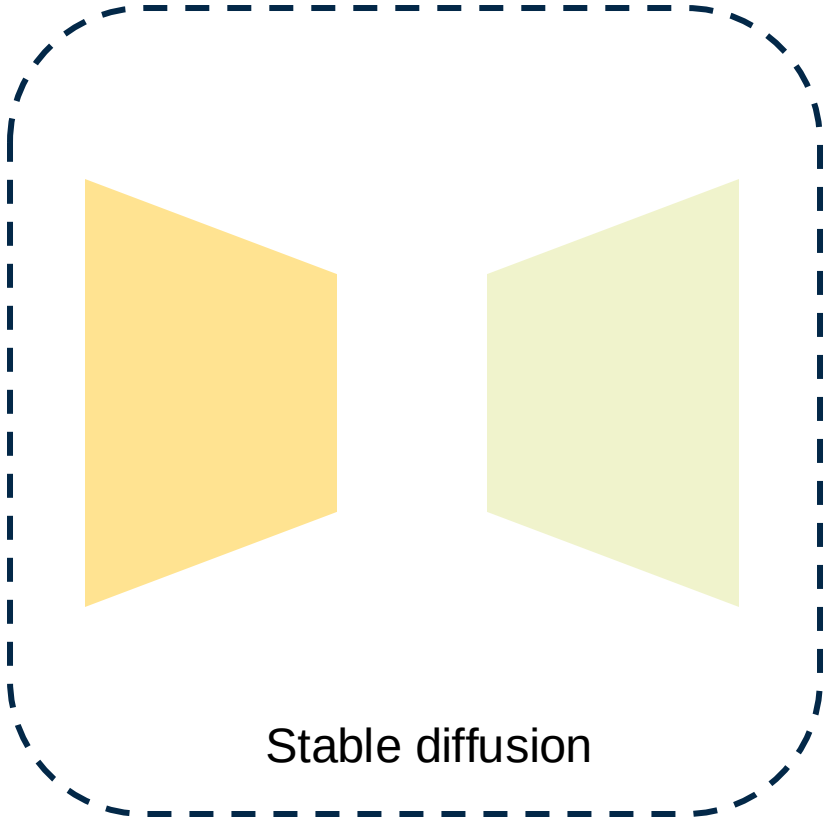


Image input:

Remove the people

Text input:



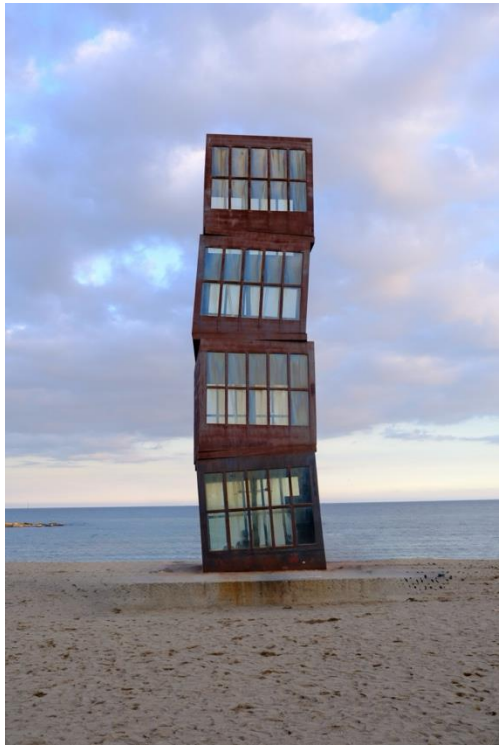


Image output:

Motivation:

- Problem: Stable diffusion will also change the content we interested

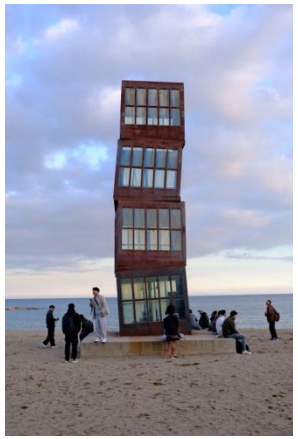


Image input:

Remove the people

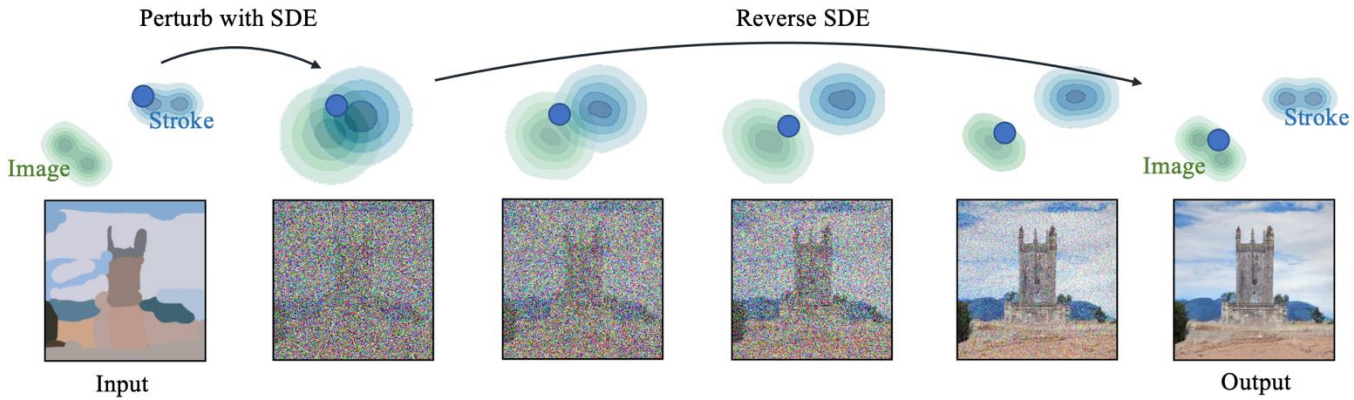
Text input:

Stable diffusion

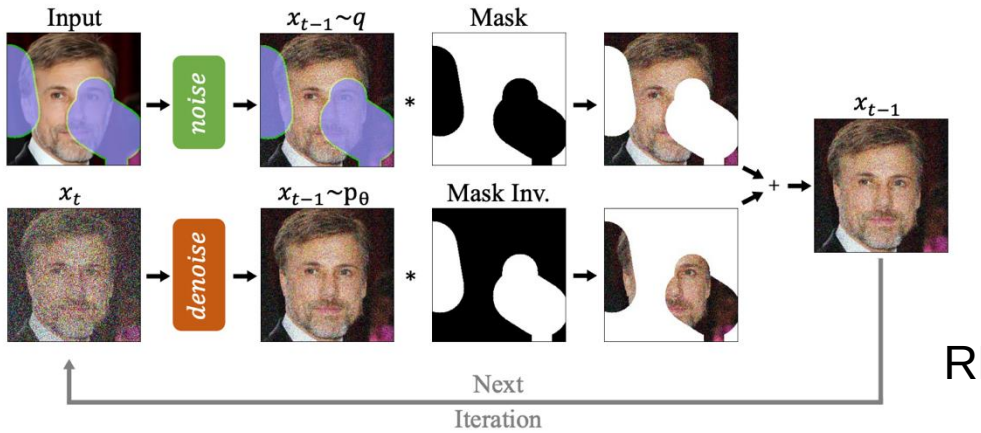
Image output:

Motivation:

- Stable Diffusion combined with other noise/modules to make it work



SD-EDIT[4]



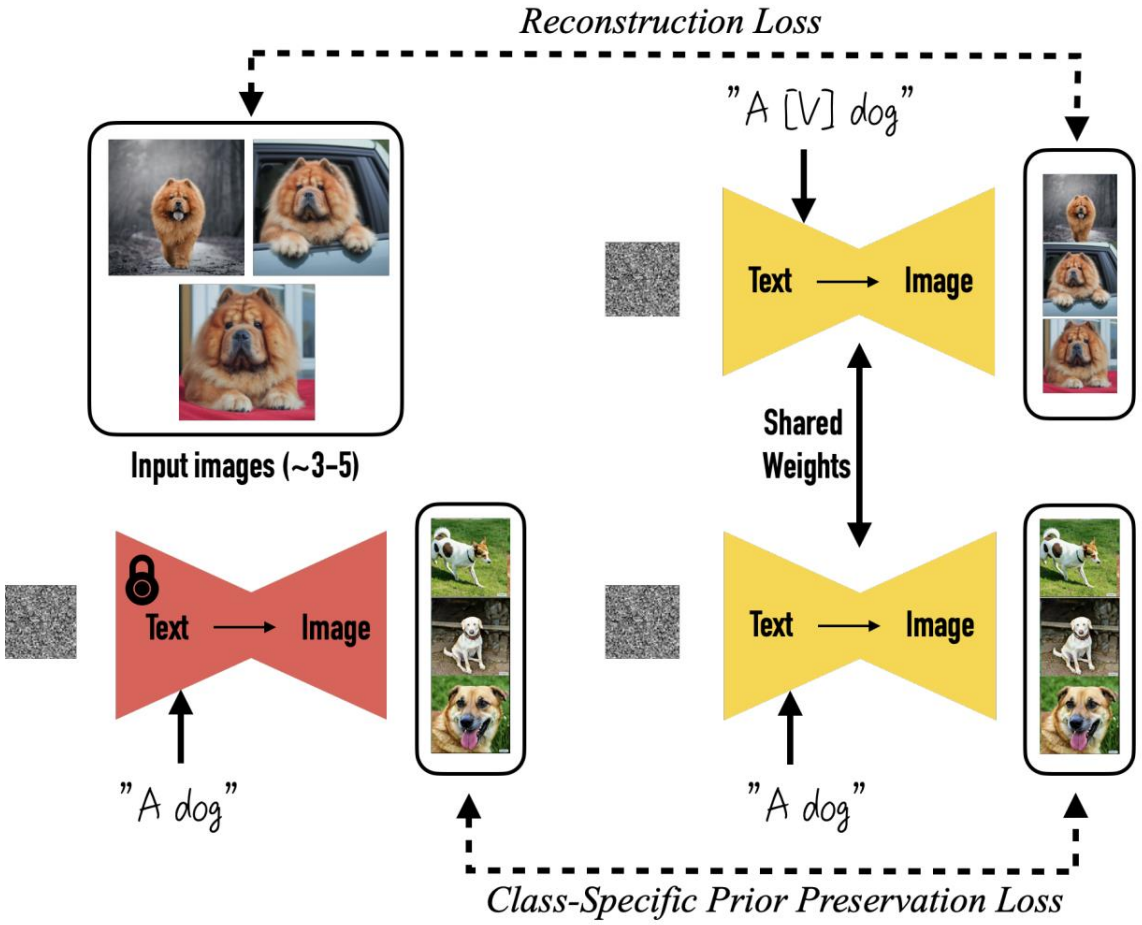
REPaint[5]

Limitation:
- Poor performance
- Extra work on masking

[4] Meng, Chenlin, et al. "Sdedit: Guided image synthesis and editing with stochastic differential equations." *arXiv preprint arXiv:2108.01073* (2021).
[5] Lugmayr, Andreas, et al. "Repaint: Inpainting using denoising diffusion probabilistic models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Motivation:

- Stable Diffusion combined with other modules to make it work

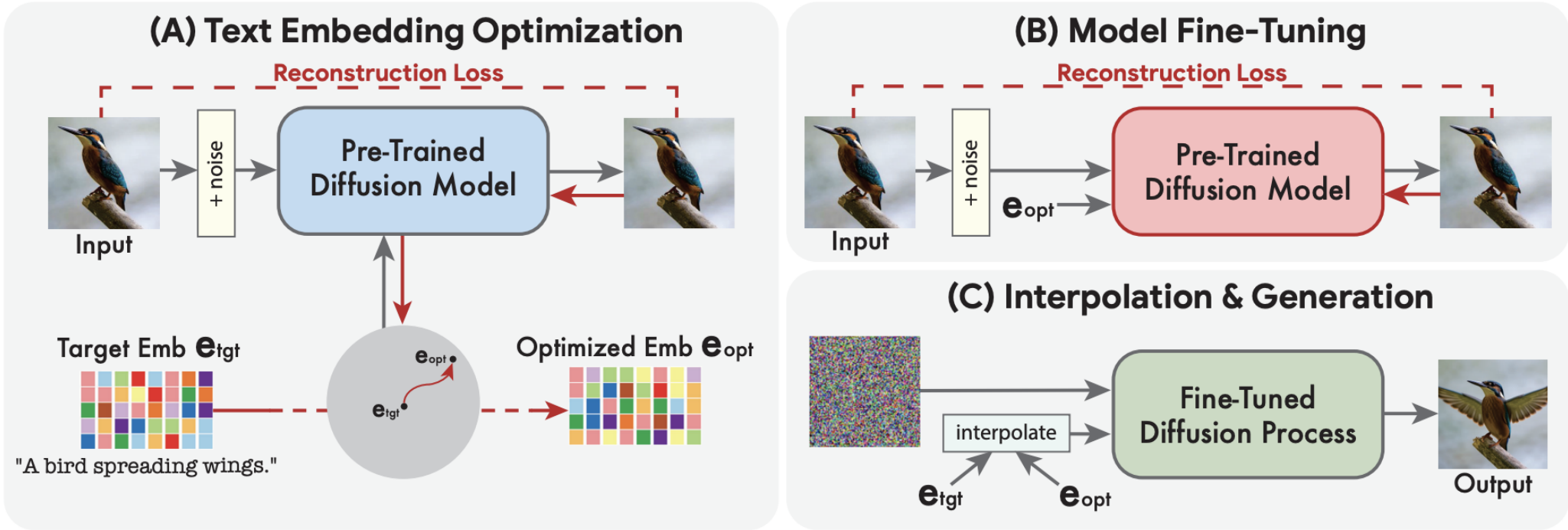


DreamBooth[6]

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

Motivation:

- Stable Diffusion combined with other modules to make it work



Imagic[7]

Limitation:

- Need to finetune whole stable diffusion model

[7] Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Motivation:

- Target: Can we train a diffusion model to follow instructions simple as below?

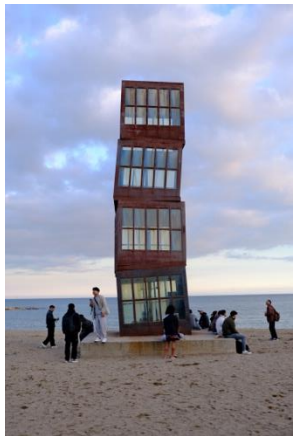
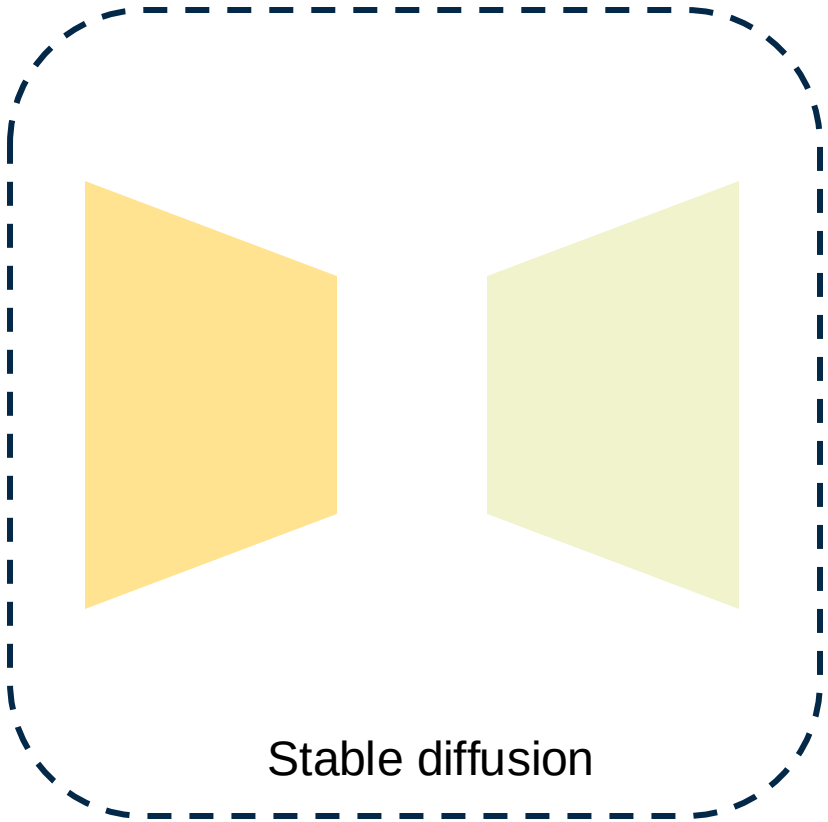


Image input:

Remove the people

Text input:



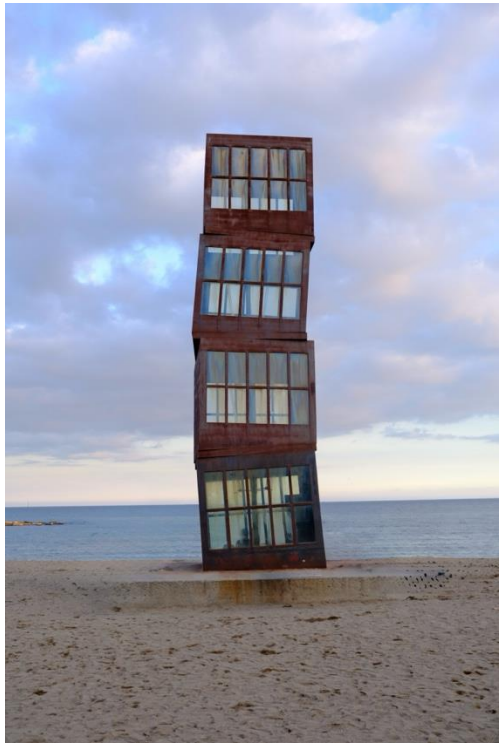


Image output:

Motivation:

- Target: Can we train a diffusion model to follow instructions simple as below?

Remove the people

Text input:

Stable diffusion

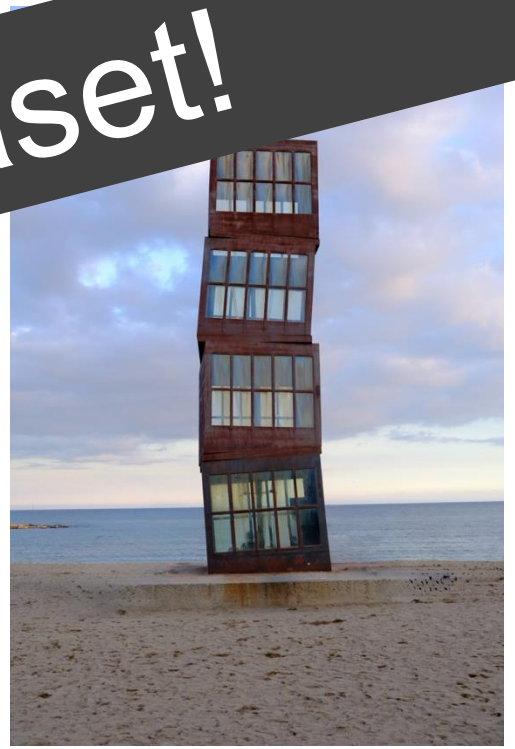


Image output:

Yes, if we have such dataset!

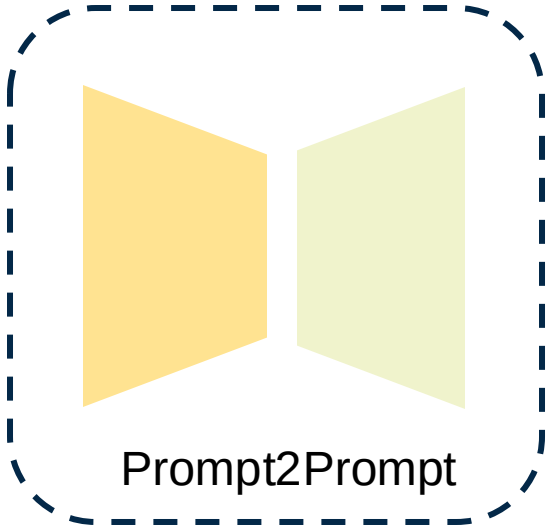
Method:

Prompt-to-Prompt[8]

- Given two similar textual descriptions as input, can synthesize image pairs while keep original structure and content

1) Photo of a cat riding on a bicycle

2) Photo of a cat riding on a car



Input text pair:

[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

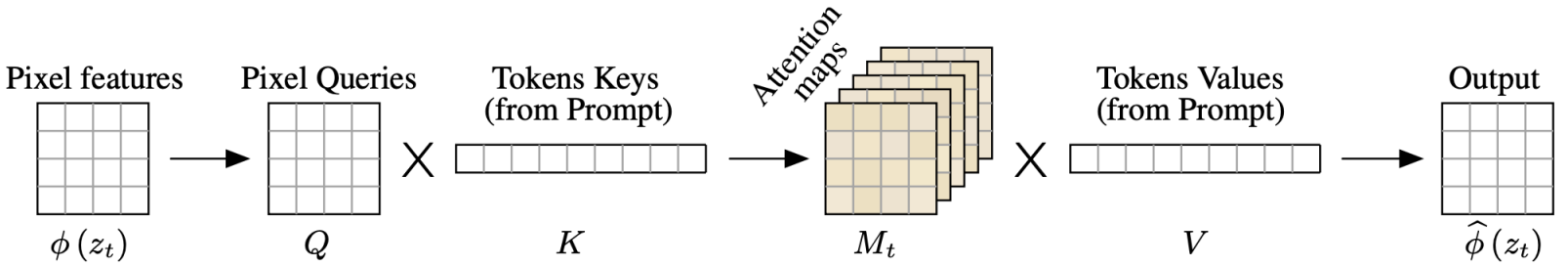
Prompt-to-Prompt[8]

- How ?

Cross attention control!

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y).$$

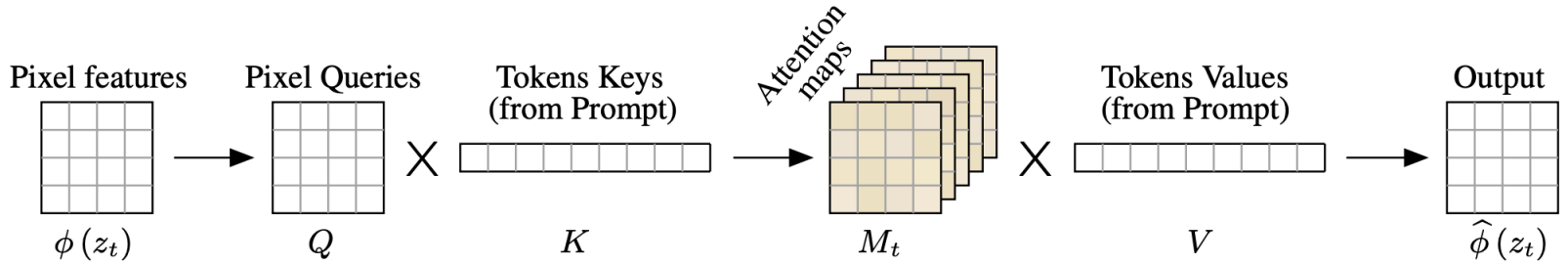
$$M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right),$$

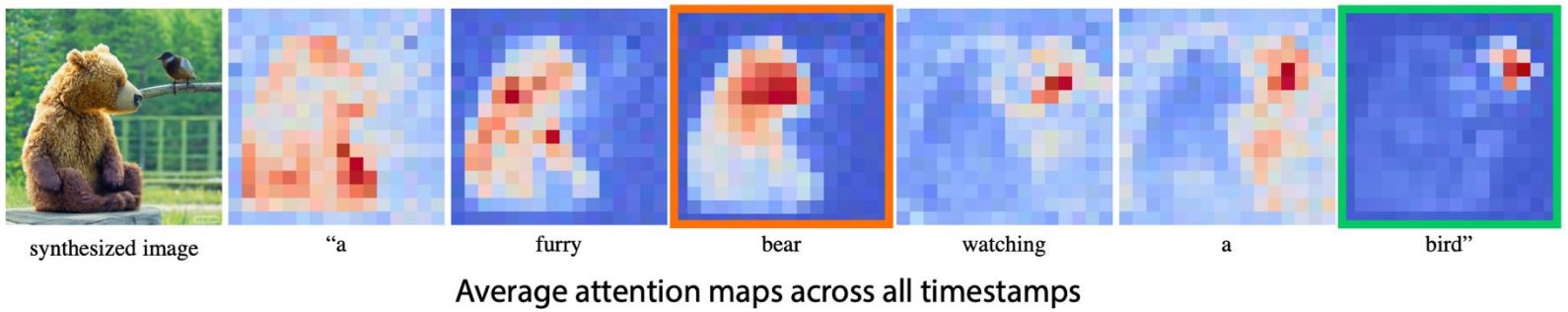


[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Prompt-to-Prompt[8]

- Cross Attention: pixels are more attracted to the words that describe them

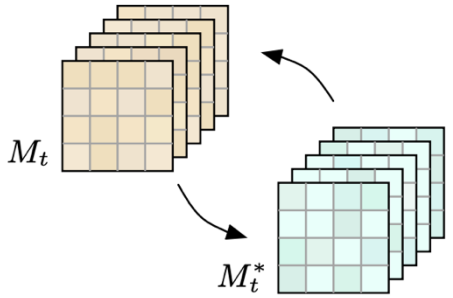




[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Prompt-to-Prompt[8]

- How to control cross attention?
  - Soucre prompt: Photo of a cat riding on a bicycle
  - Target prompt: Photo of a cat riding on a car

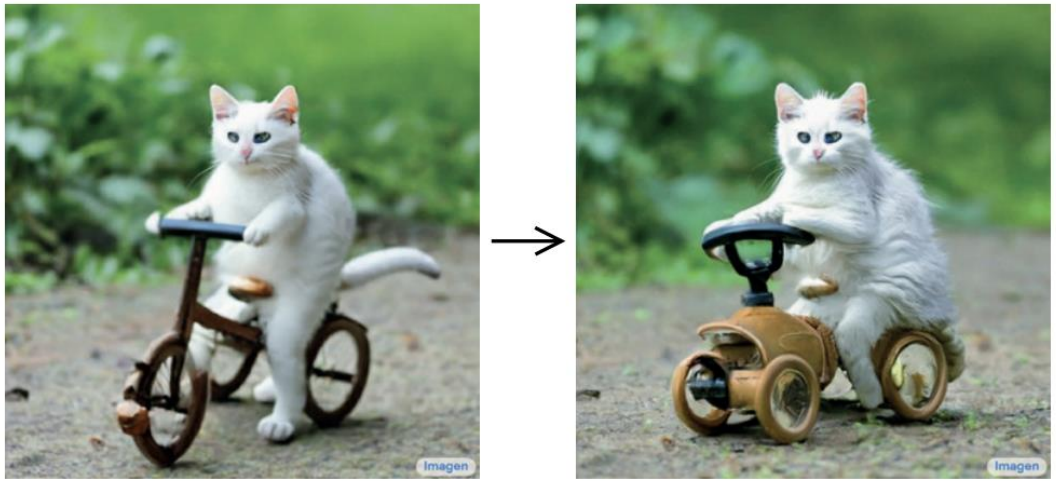


Word Swap

- The Attention map will be different!

Algorithm 1: Prompt-to-Prompt image editing

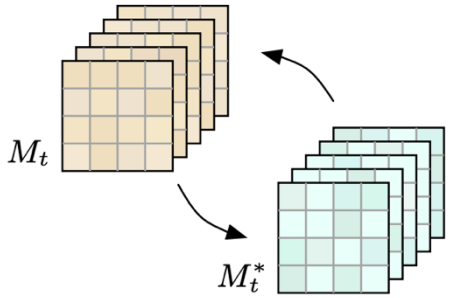
Input: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$, and a random seed s.
Output: A source image x_{src} and an edited image x_{dst}.
$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed s;
$z_T^* \leftarrow z_T$;
for $t = T, T-1, \ldots, 1$ **do**
  $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$;
  $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$;
  $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$;
  $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$;
end
Return (z_0, z_0^*)



[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Prompt-to-Prompt[8]

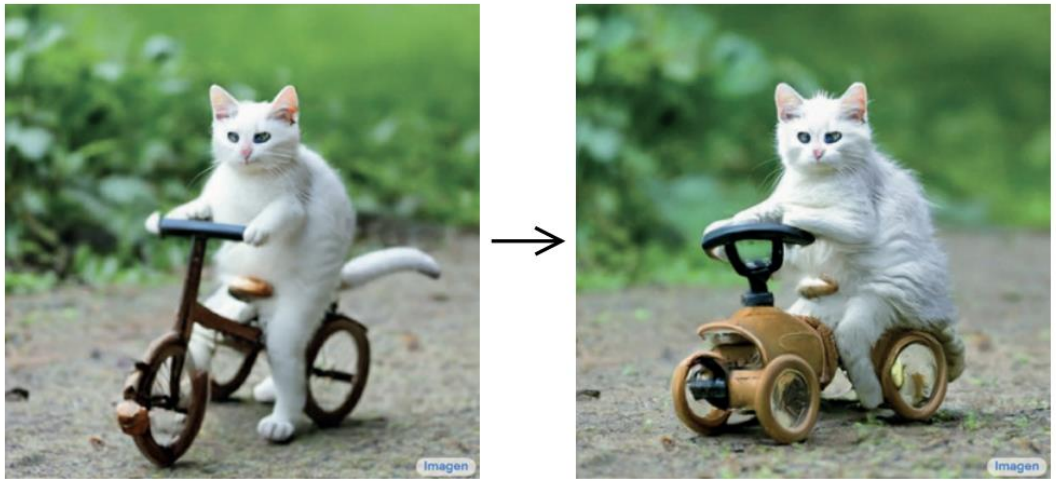
- How to control cross attention?
  - Soucre prompt: Photo of a cat riding on a bicycle
  - Target prompt: Photo of a cat riding on a car



Word Swap

- Scheduling is also important

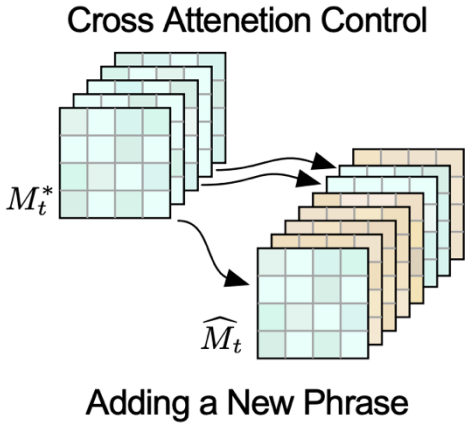
$$Edit(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise.} \end{cases}$$

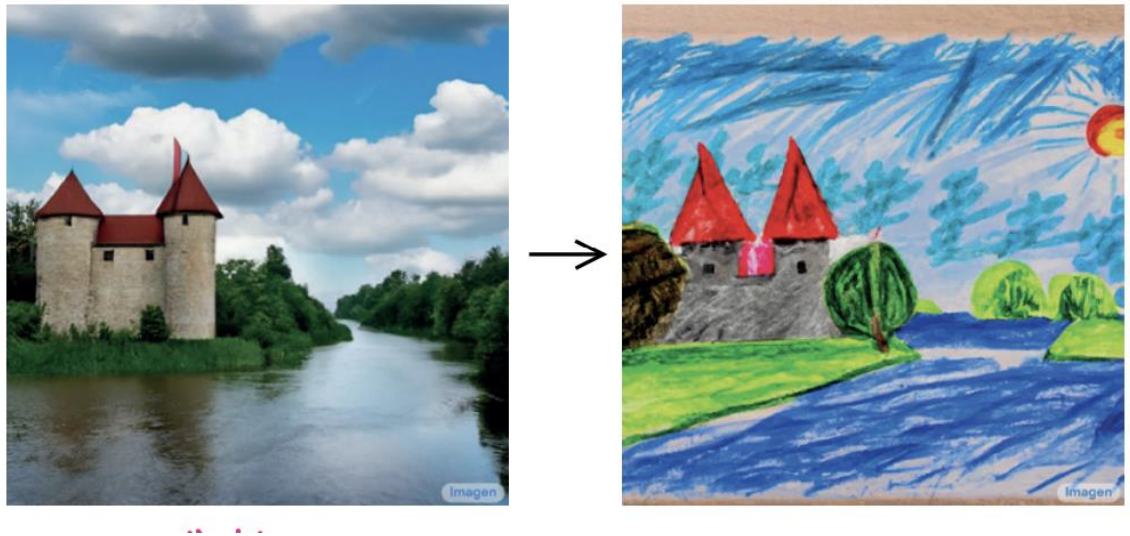


[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Prompt-to-Prompt[8]

- How to control cross attention?
  - Soucre prompt: a castle next to a river
  - Target prompt: children drawing of a castle next to a river

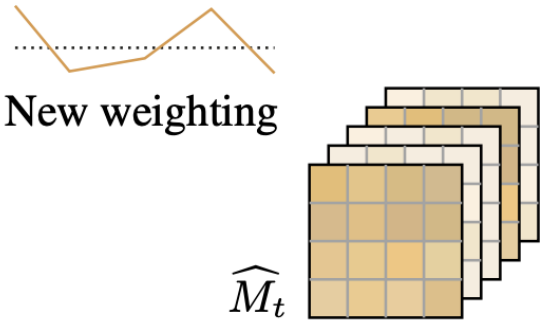




[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

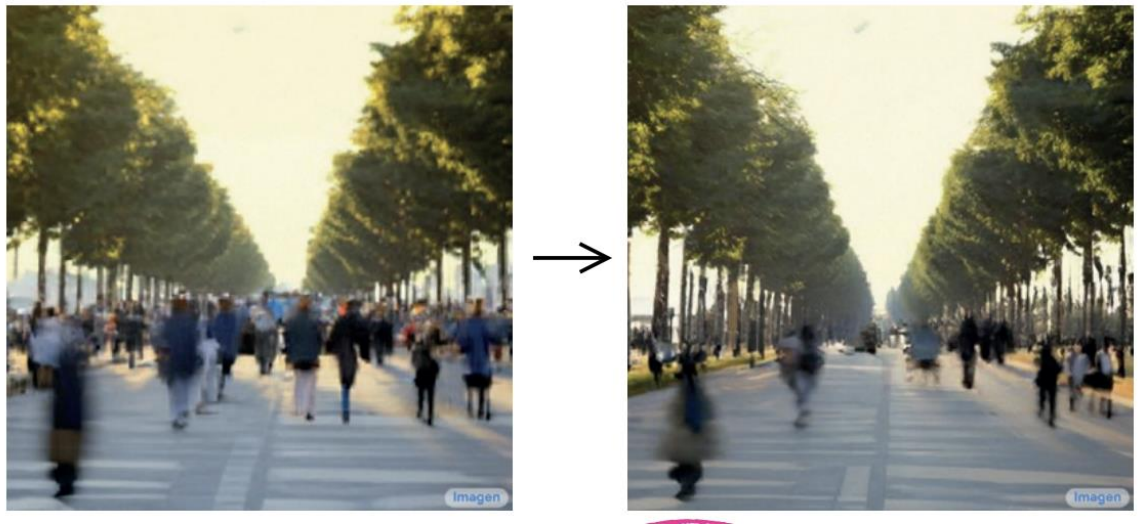
Prompt-to-Prompt[8]

- How to control cross attention?
  - Soucre prompt: The boulevards are crowded today
  - Target prompt: The boulevards are crowded today



Attention Re–weighting

- The Attention map will be different!

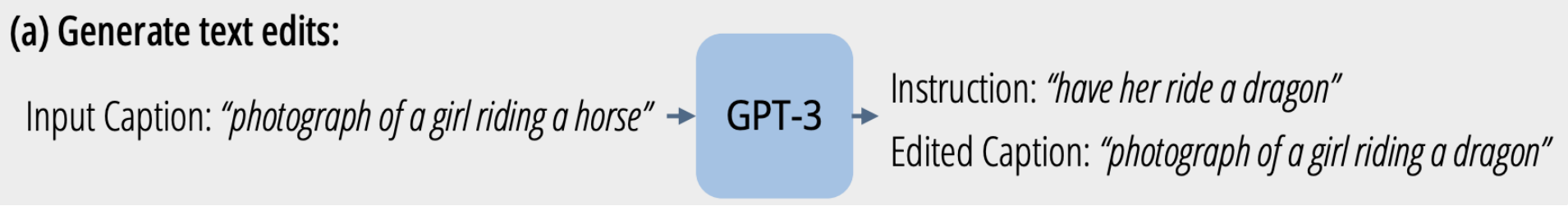


[8] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

GPT-3[9]

- With Prompt to Prompt
  - We can generate images pair given captions pair

- How to automatically get captions pair?
  - Large pretrained language model GPT-3
  - Finetune GPT-3 on 700 captions and manually wrote instructions and output captions



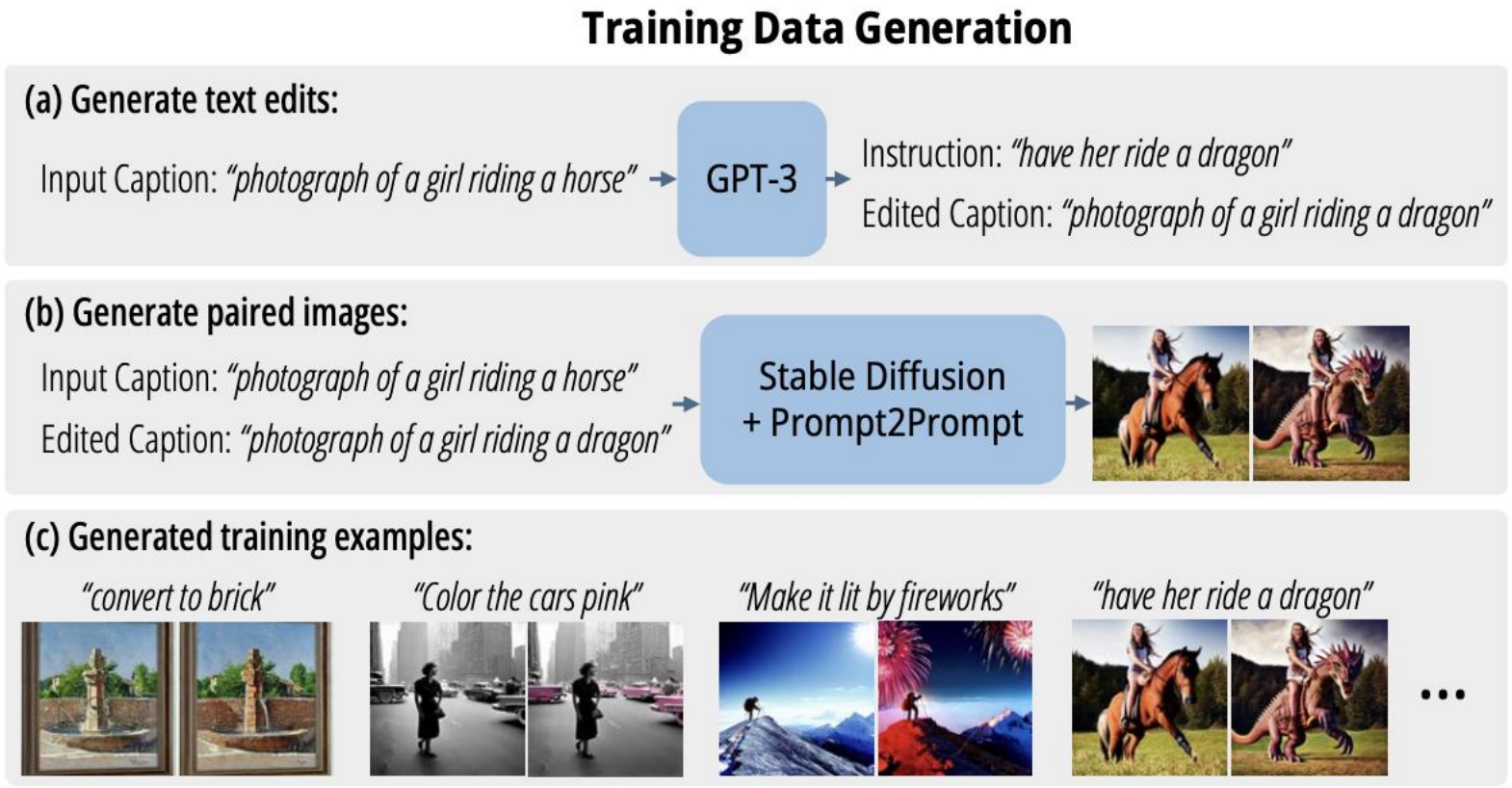
[9] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

GPT-3[9]

- How to automatically get captions pair?
  - Large pretrained language model GPT-3
  - Finetune GPT-3 on 700 captions and manually wrote instructions and output captions

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 generated (>450,000 edits)	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

Dataset Generate Flow



(a) Without Prompt-to-Prompt. (b) With Prompt-to-Prompt.

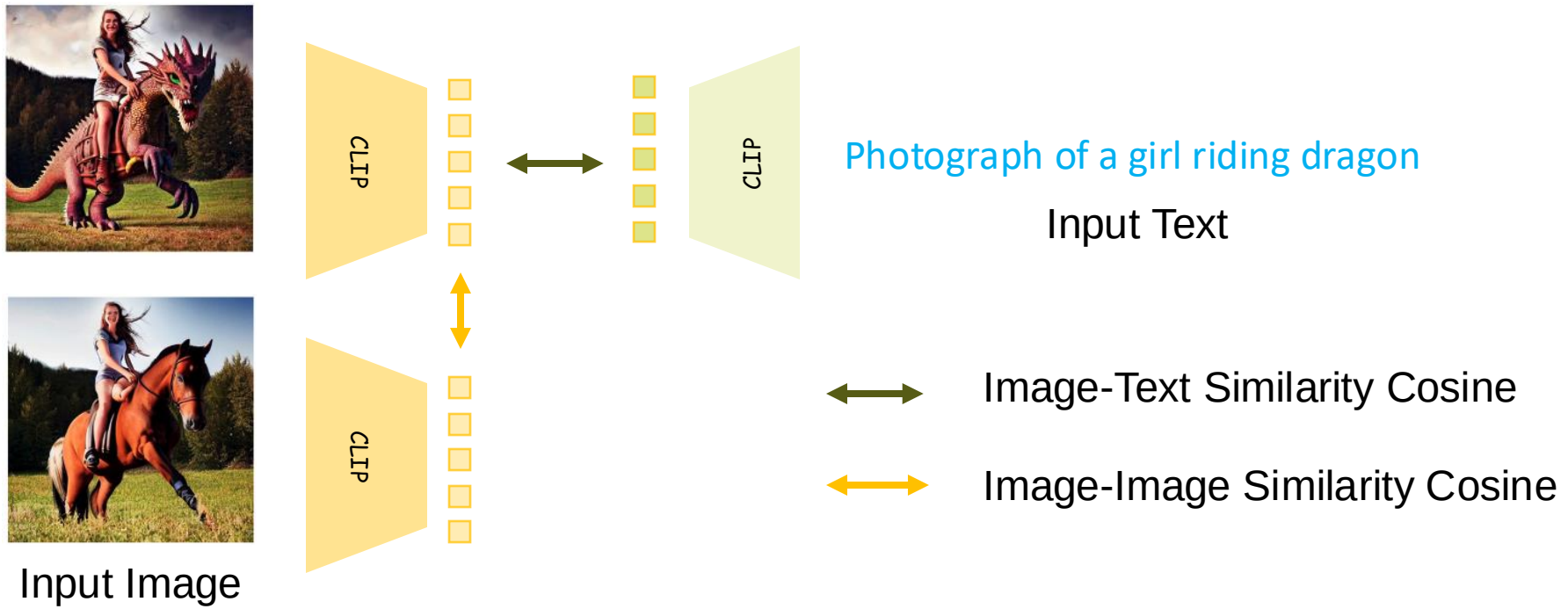
Importance of using Prompt-to-Prompt

Dataset Quality Control

- What if Prompt-to-Prompt fail?

Dataset Quality Control

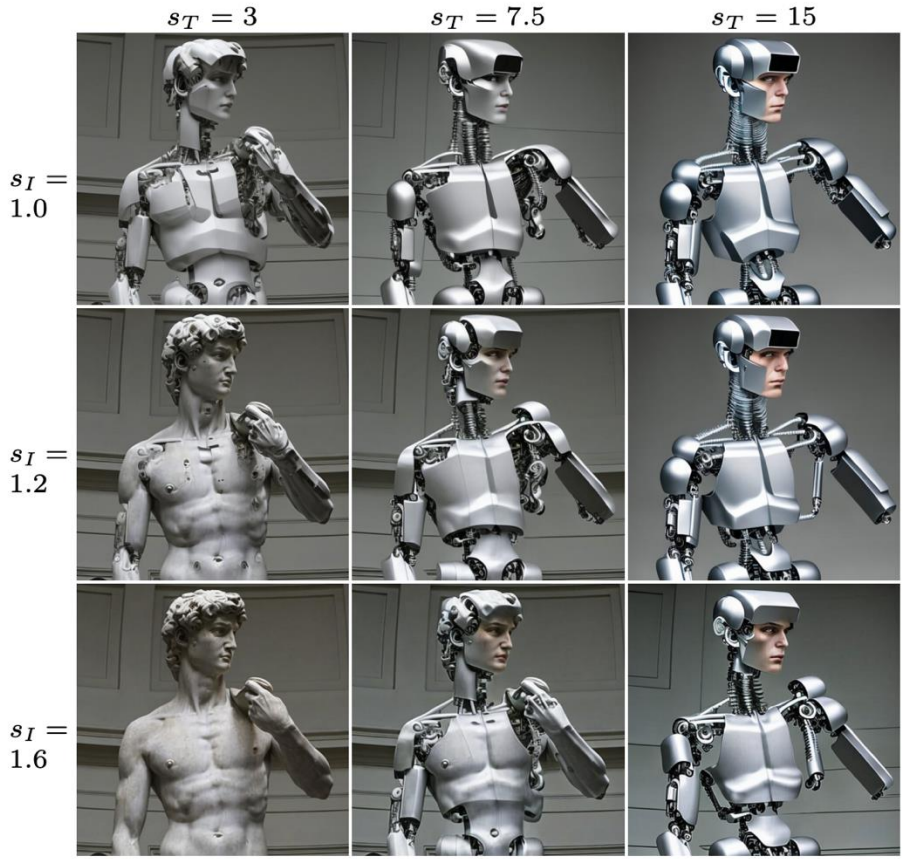
- How to control the data quality?
  - Try different prompt to prompt hyperparameters
    - One caption -> 100 images pair
  - Filter the pairs with CLIP[10]
    - CLIP can estimate image-text similarity



[10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Classifier Free Guidance

- Stable diffusion training in both conditional and unconditional
  - 5% randomly set to null during training
  - During inference:



Edit instruction: *"Turn him into a cyborg!"*

c_I Image conditioning

c_T Text instruction conditioning

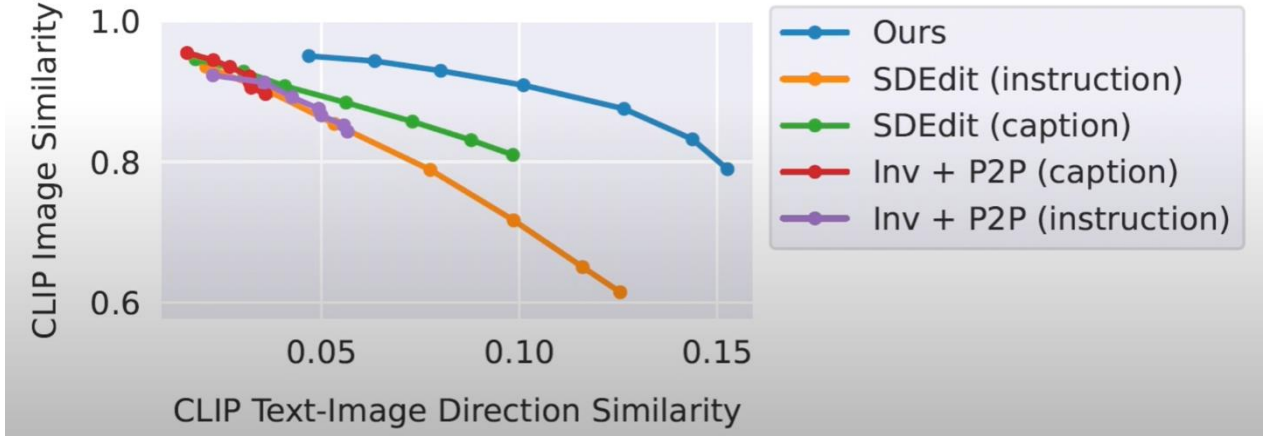
s_I s_T Guidance Scales

$$\tilde{e}_\theta(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))$$

Experiments:

Experiment:

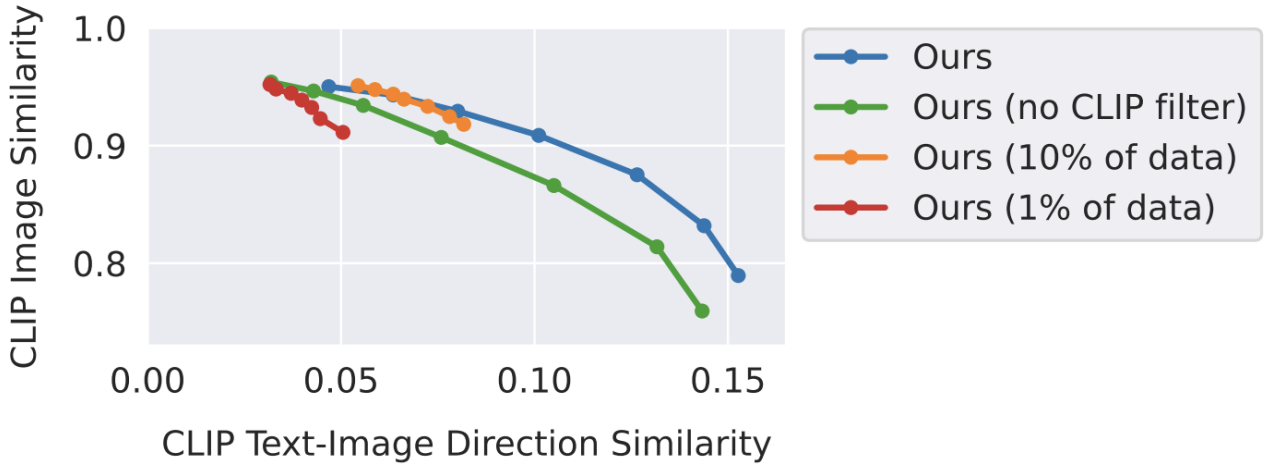
- Quantitative Results:
  - Compare With baseline SDEdit, P2P
  - CLIP Text-image Direction Similarity:
    - How much the change in text captions agrees with the change in the images
  - Fix text guidance s_T =7.5 and vary the image guidance s_I =[1.0, 2.2]



$$\Delta T = E_T\left(t_{target}\right) - E_T\left(t_{source}\right) \ ,$$
$$\Delta I = E_I\left(G_{train}\left(w\right)\right) - E_I\left(G_{frozen}\left(w\right)\right) \ ,$$
$$\mathcal{L}_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I|\,|\Delta T|} \ .$$

Experiment:

- Quantitative Results:
  - Importance of scale of dataset and filtering
  - Fix text guidance s_T =7.5 and vary the image guidance s_I =[1.0, 2.2]



$$\Delta T = E_T\left(t_{target}\right) - E_T\left(t_{source}\right) \ ,$$
$$\Delta I = E_I\left(G_{train}\left(w\right)\right) - E_I\left(G_{frozen}\left(w\right)\right) \ ,$$
$$\mathcal{L}_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I|\,|\Delta T|} \ .$$

Experiment:

- Qualitative Results:
  - Trained on generated images, but generalize to real images and text instructions

Input

"Make it a grocery store"

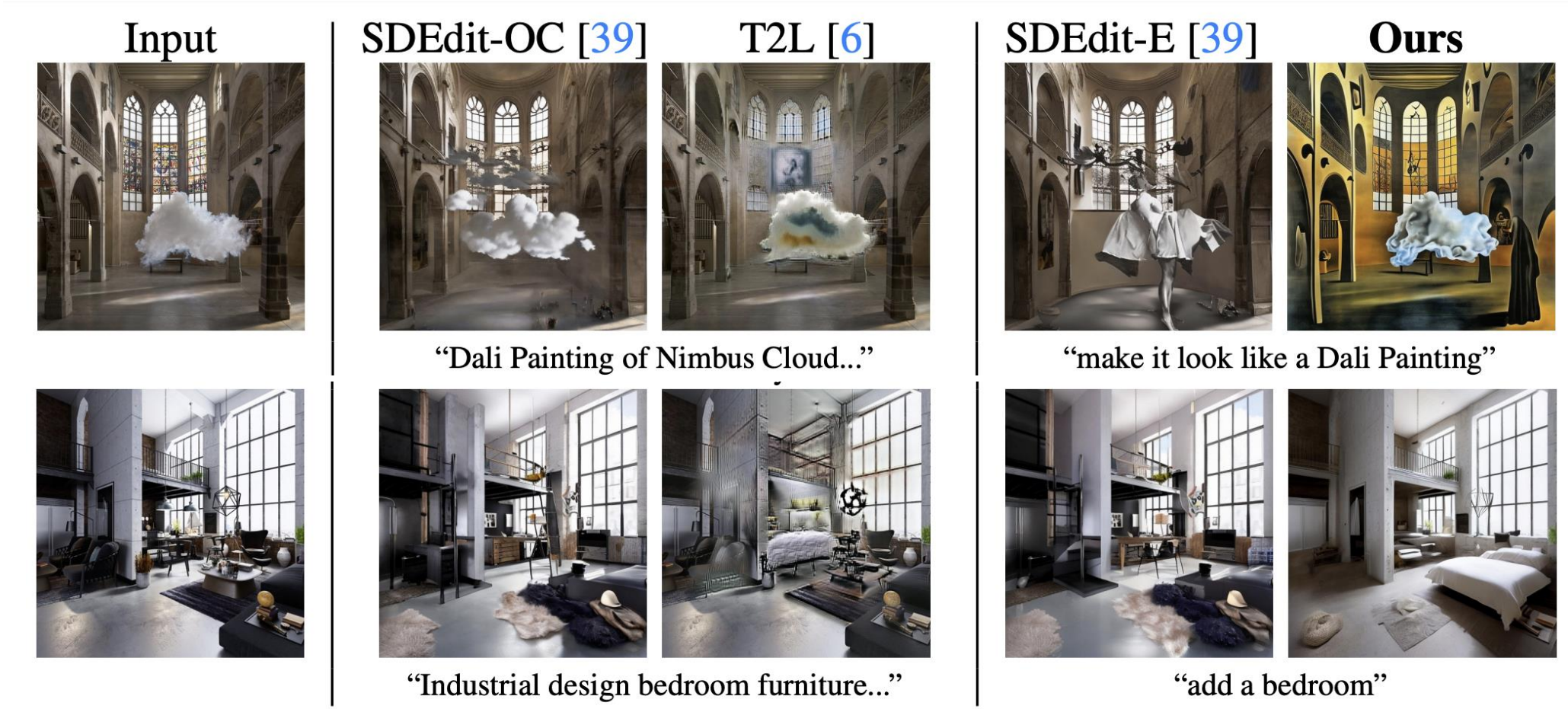
Input

"Add boats on the water"

"Replace the mountains with a city skyline"

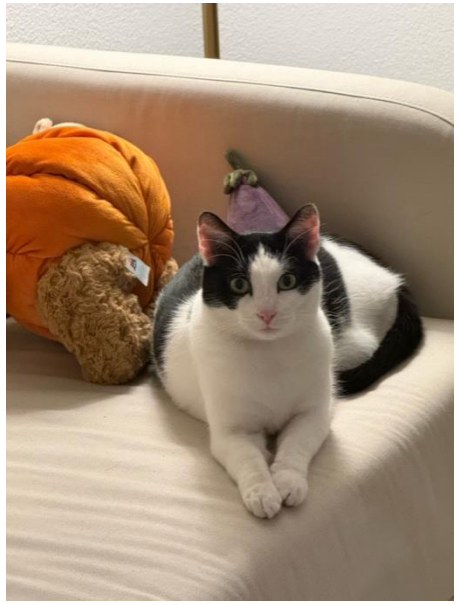
Experiment:

- Qualitative Results:
  - Compared with baseline



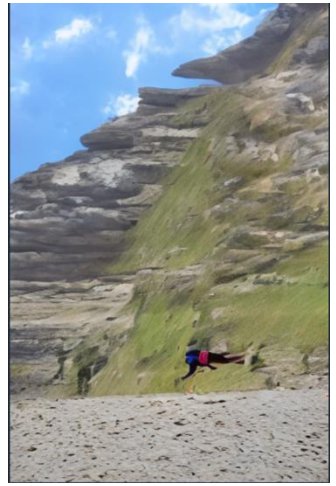
Custom demo:

Have him wear a hat.



Turn the cat to dog

Remove the people



Make it Minecraft

Some Thoughts:

Contribution:

- Large Scale Dataset
  - 454,445 pairs after filtering

- Fast and simple text instruction based method
  - Classifier Free guidance on two conditionings
  - Impressive results and generalization to real / drawing

- Inspiring:
  - Take advantage of pretrained weights

Limitation:

- Large Scale Dataset, large enough?
  - 0.4 million vs 3 billion (Imagen) vs 5.8 billion (stable diffusion)

- Large Scale Dataset, good enough?
  - Limited data quality control with CLIP,
  - 0.75 threshold image to image, 0.2 threshold image to text
  - User Study

CLIP Score:
0.96

0.33 photograph of a girl riding a dragon

0.21 photograph of a girl riding a Horse

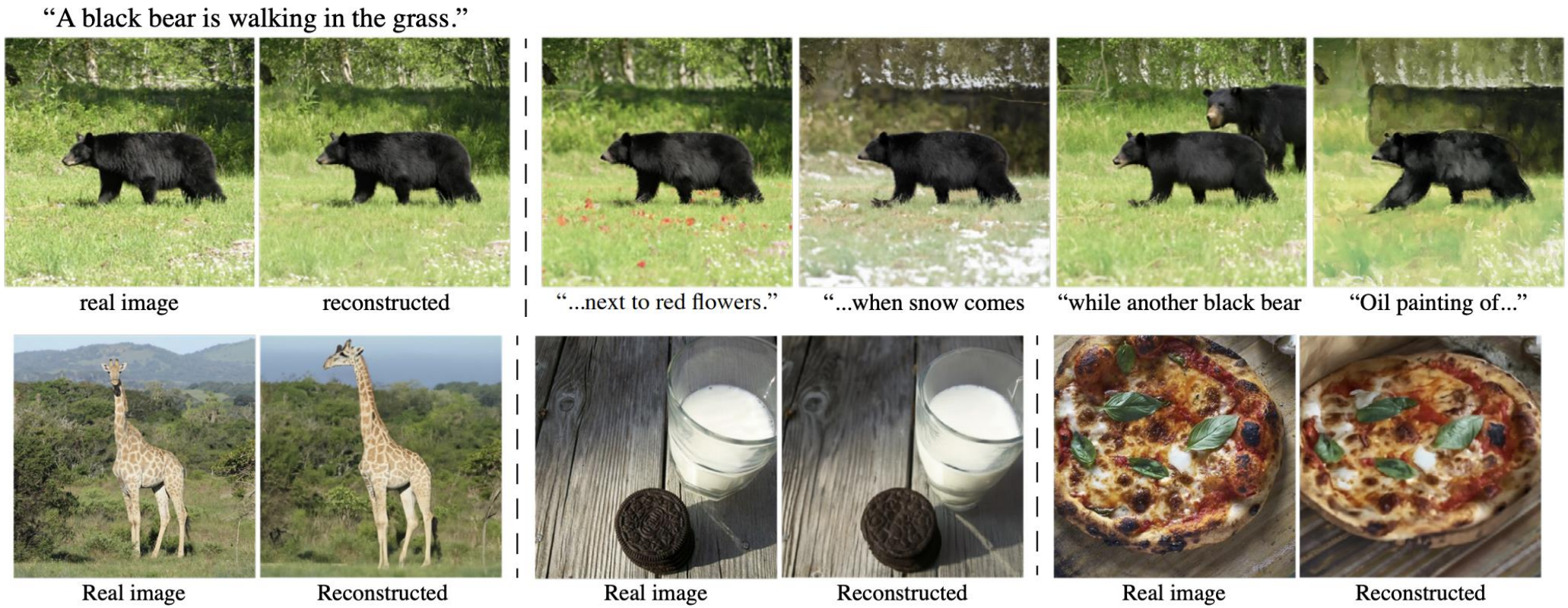
0.14 photograph of a girl riding a *bicycle*

0.34 photograph of a dragon ***riding*** a girl

0.32 photograph of a dragon ***chasing*** a girl

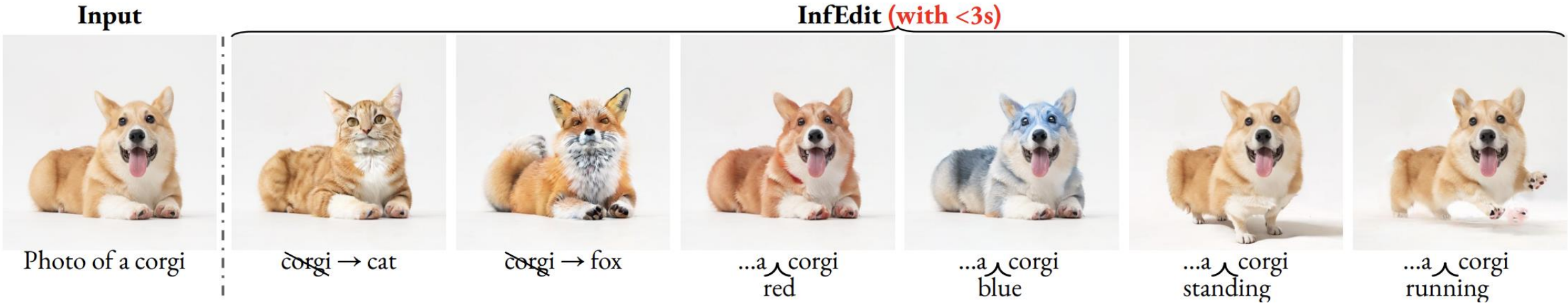
Potential Improvement:

- Data Scalability
  - Need to train on synthetic image, can it train on real image?
  - Need to inverse the Gaussian process!



Potential Improvement:

- Data Scalability
  - Improved work on this: InfEdit[11]



  - Improved work on captions: ChatGPT, DeepSeek…

[11] Xu, S., Huang, Y., Pan, J., Ma, Z., & Chai, J. (2024). Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9452-9461)..

Q&A:

Q&A: