



MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing

Mingdeng Cao^{1,2*} Xintao Wang²✉ Zhongang Qi² Ying Shan² Xiaohu Qie² Yinqiang Zheng¹✉

¹The University of Tokyo

²ARC Lab, Tencent PCG

ICCV 2023



Introduction

Abstract



Input real image

“... jumping ...”



“A sitting boy” → “... standing ...”



Input real image

“... giving a thumbs up ...”



“Elon Musk → ... side view ...”



“An apple” → “... two ...”



“A standing bird” → “... spreading wings ...”

MasaCtrl can perform text-based non-rigid image synthesis and real image editing without finetuning.

Contributions

- 1) A **tuning-free** method to achieve consistent image synthesis and complex image editing.
- 2) An effective **mutual self-attention** mechanism.
- 3) A **masked-guided** mutual self-attention, where the mask can be easily computed from the cross-attentions.

The effectiveness of our proposed MasaCtrl in both **consistent image generation** and **complex non-rigid real image editing**.



"...sitting..."

"...laying..."

"...laying..." (Ours)



Source

w/o mask guidance

with mask guidance

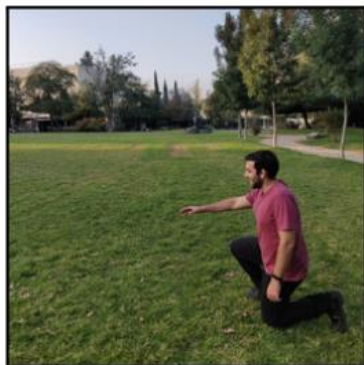


Related Work

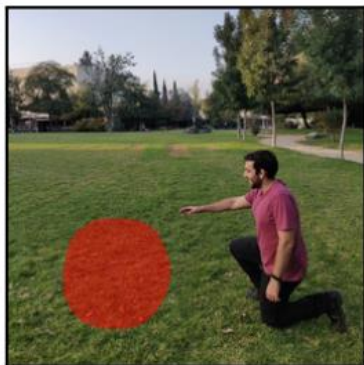
Text-guided Image Editing

Blended latent diffusion

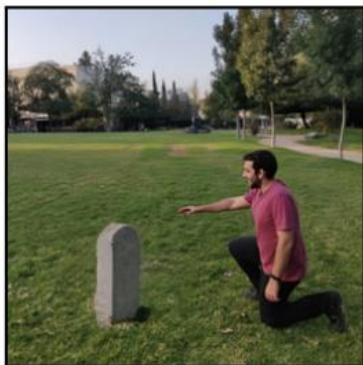
Require extra masks to edit local regions of the image;



Input image



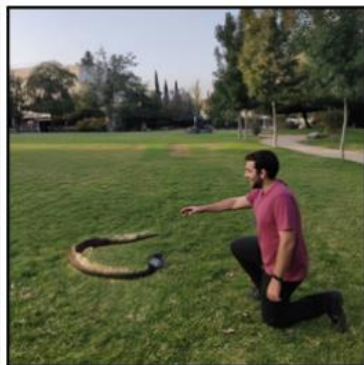
Input mask



“gravestone”



“toy truck”



“snake”

Text-guided Image Editing



DiffusionCLIP

Can edit global aspects of the image by changing the text prompt directly, but **cannot modify local details**;



Text-guided Image Editing

Prompt-to-prompt

use **cross-attention** or to edit both global and local aspects of the image by changing the text prompt directly, but they tend to preserve the original layout of the source image and **fail to handle non-rigid transformations**.



“The boulevards are crowded today.”



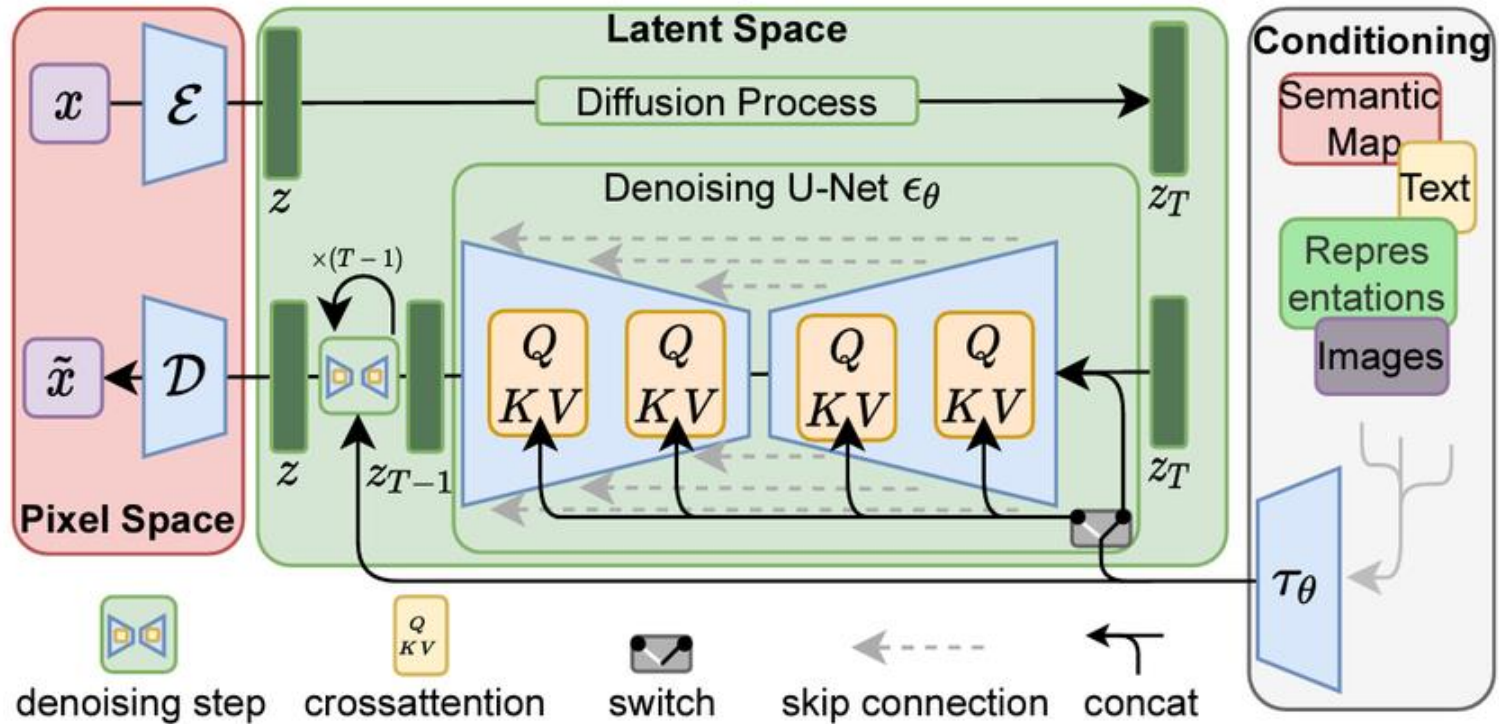
“Photo of a cat riding on a ~~bicycle~~.”

car

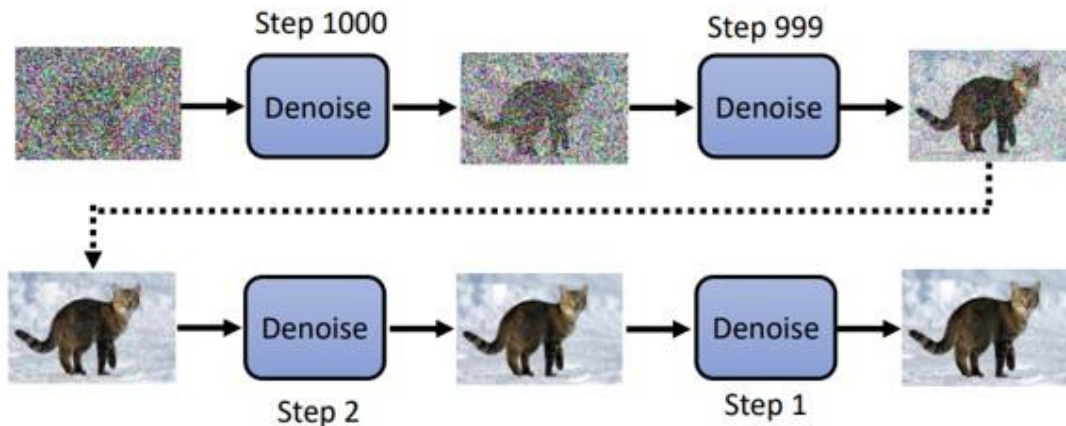
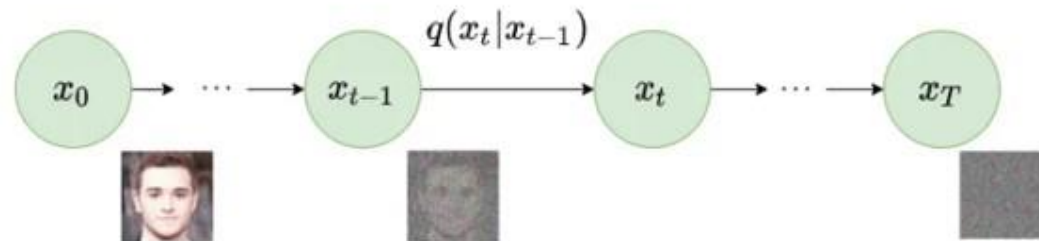


Preliminaries

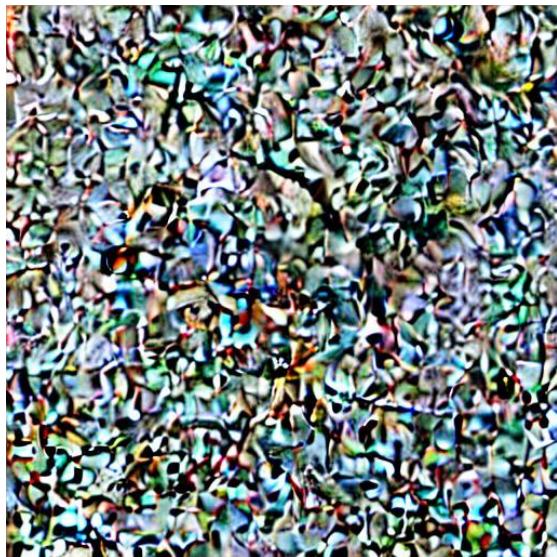
Stable Diffusion



Diffusion Model



DDIM Inversion

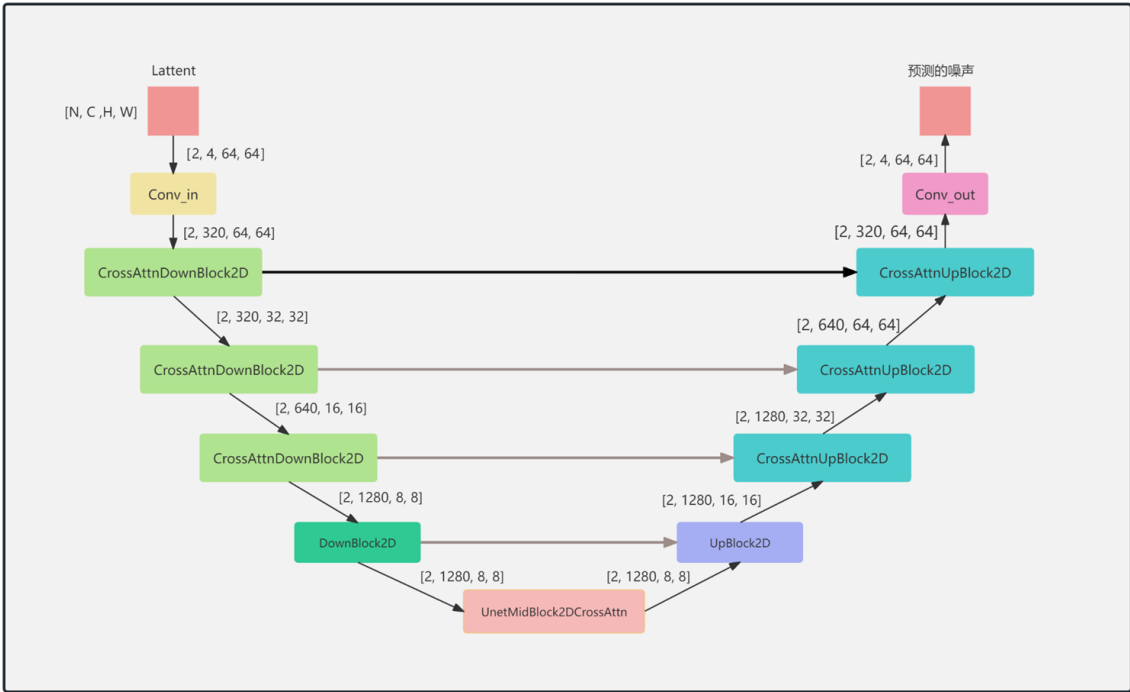


invert

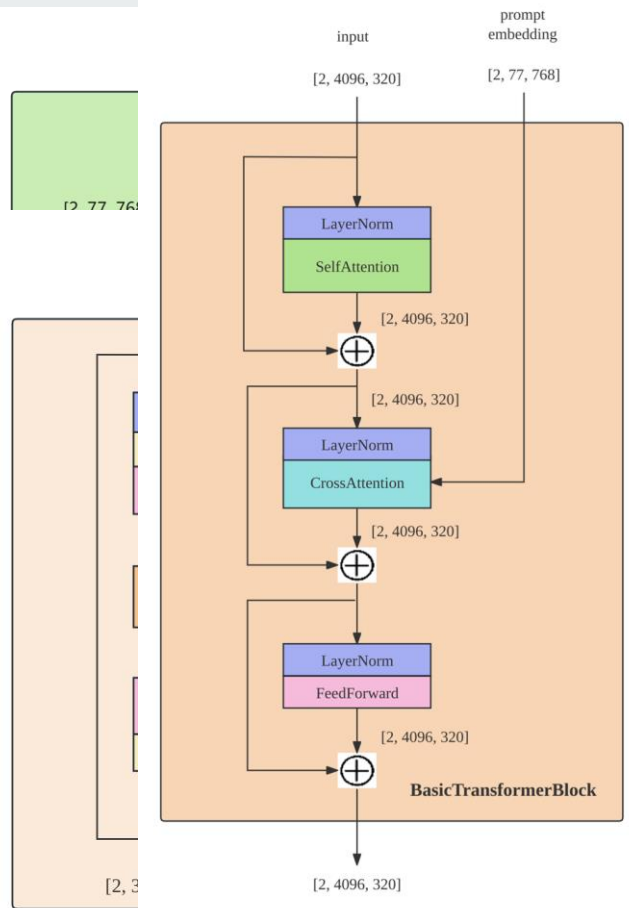


reconstruct

Attention Block



CSDN @wu_jiacheng



CSDN @wu_jiacheng

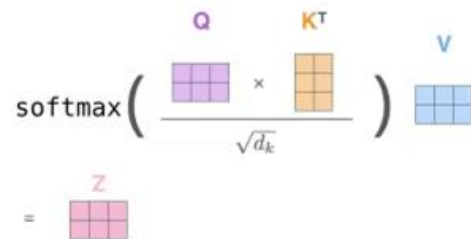
Attention Mechanism

Attention is to map the query and key into the same high-dimensional space to **calculate the similarity**.

$$X \times W^Q = Q$$


$$X \times W^K = K$$


$$X \times W^V = V$$


$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V = Z$$


Attention in Stable Diffusion

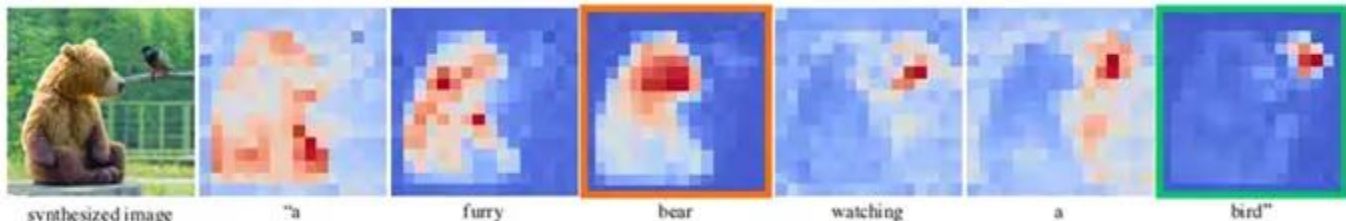
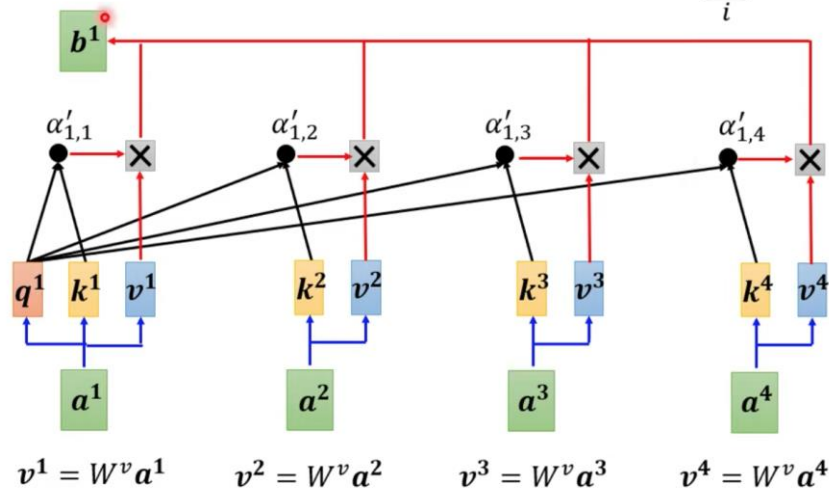
At denoising step t , the features from the previous $(l-1)$ -th basic block first pass through the residual block to generate intermediate **features** f_t^l .

- Then they are reorganized by a **self-attention** layer.
- Receive textual information from the given text prompt P by the following **cross-attention** layer.

Self-attention

Extract information based on attention scores

$$b^1 = \sum_i \alpha'_{1,i} v^i$$



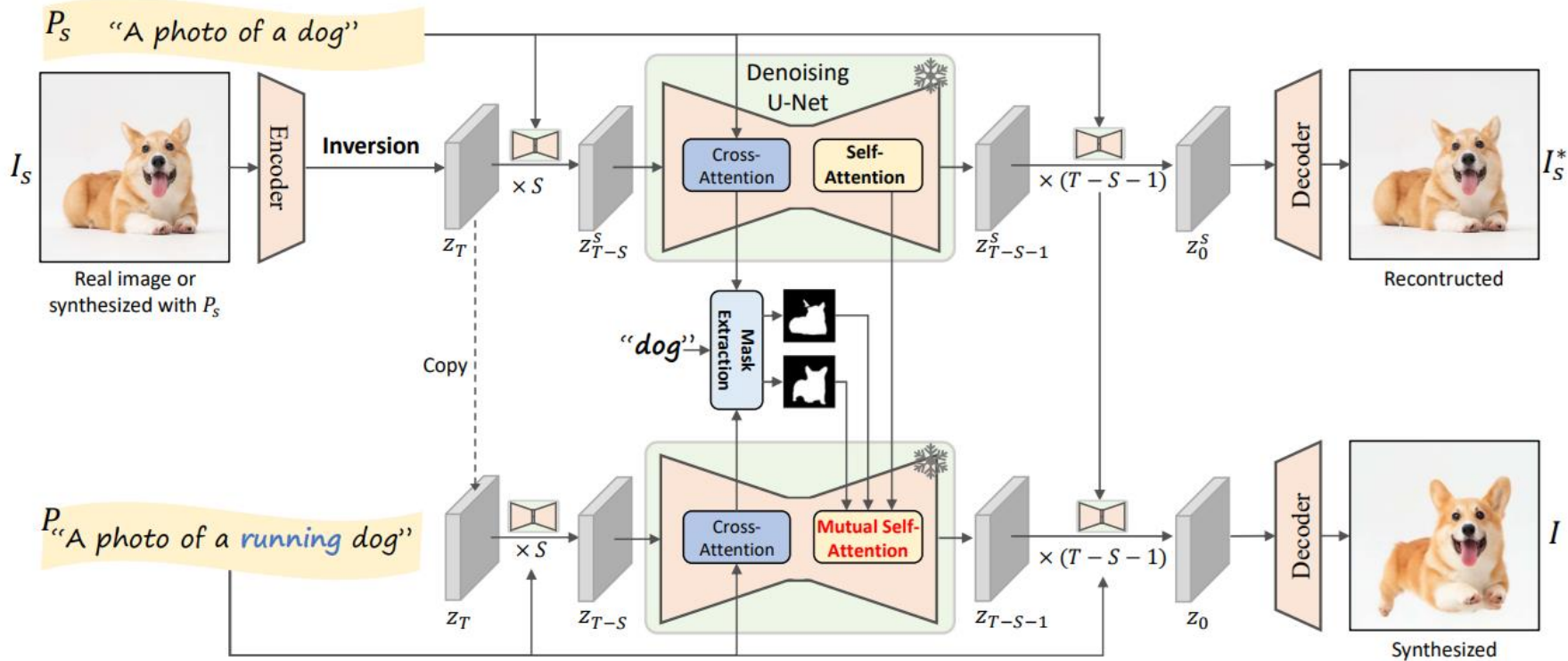
Average attention maps across all timestamps



Method

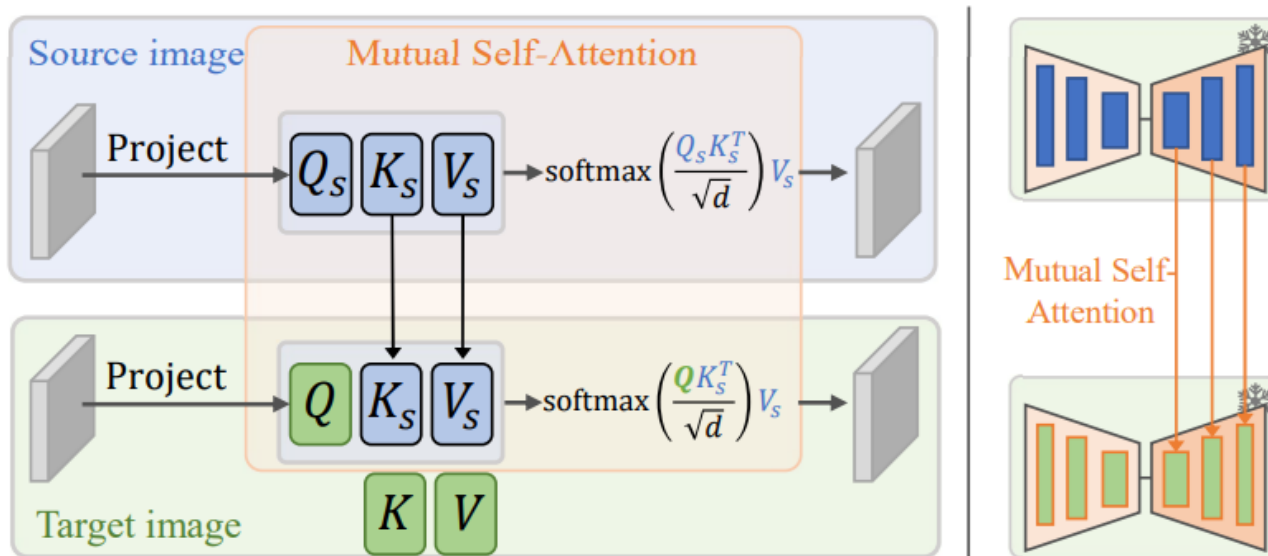
Pipeline

Input source image and target prompt P | Initial layout synthesis | Source image content querying with mutual self-attention | Output synthesized image



1. Mutual Self-Attention

They propose mutual self-attention, which converts the existing self-attention in T2I models into ‘cross-attention’, where the crossing operation happens in the self-attentions of two related diffusion processes.

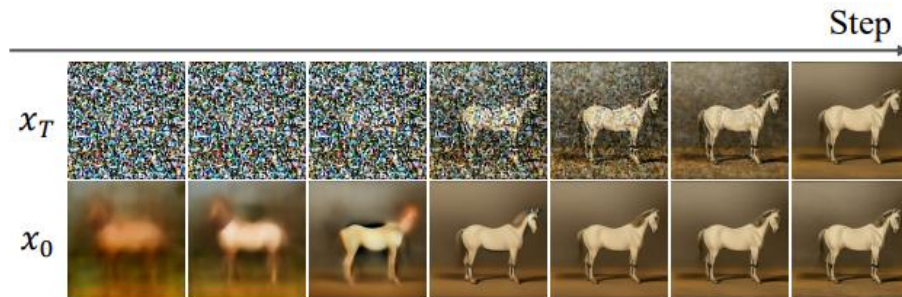


1. Mutual Self-Attention

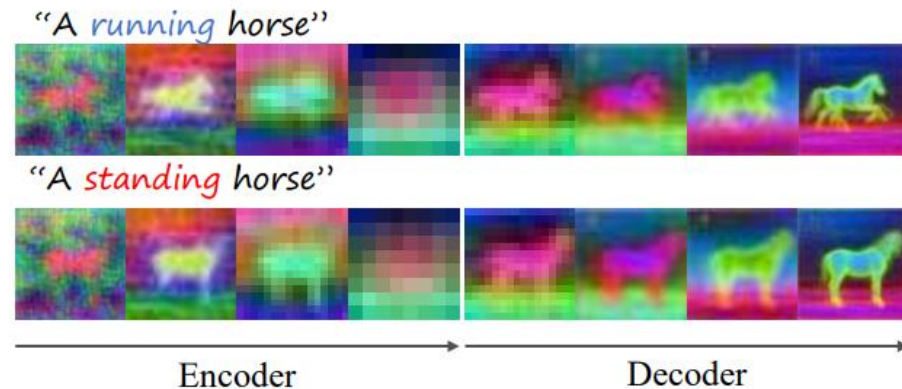
However, intuitively performing such attention control on **all layers** among **all denoising steps** will result in an image I that is nearly the same as the reconstructed image I_s .

only in the decoder part of the U-Net
after several denoising steps and layers.

$$\text{EDIT} := \begin{cases} \{Q, K_s, V_s\}, & \text{if } t > S \text{ and } l > L, \\ \{Q, K, V\}, & \text{otherwise,} \end{cases}$$



(a) Intermediate results in denoising process



(b) Query feature visualization

2. Mask-Guided Mutual Self-Attention

The cross-attention maps correlating to the prompt tokens contain most information of the shape and structure.

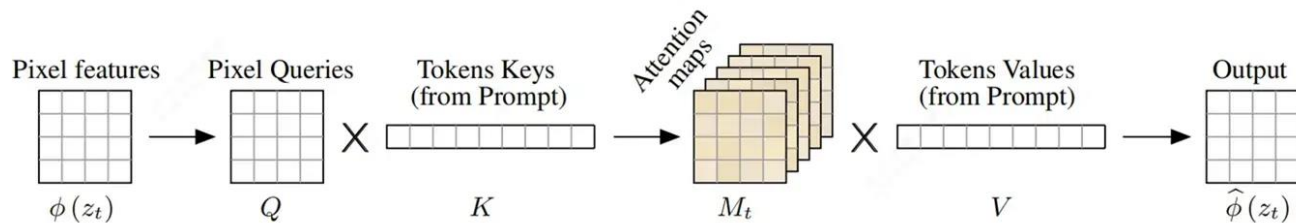
$$f_o^l = \text{Attention}(Q^l, K_s^l, V_s^l; M_s),$$

$$f_b^l = \text{Attention}(Q^l, K_s^l, V_s^l; 1 - M_s),$$

$$\bar{f}^l = f_o^l * M + f_b^l * (1 - M),$$



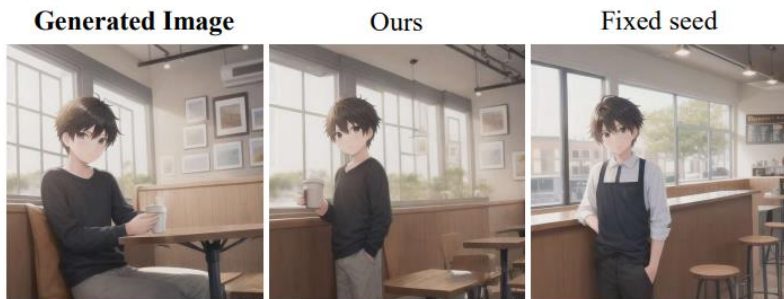
(b) Mask extraction from cross-attention maps



Text to Image Cross Attention


3. Integration to Controllable Diffusion Models

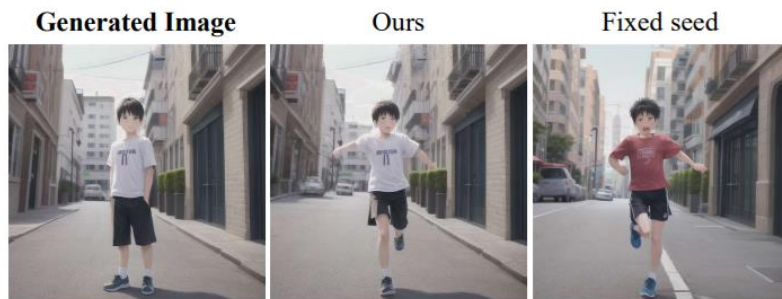
Our method can be easily integrated into existing controllable image synthesis method.



"A boy, indoors, sitting, coffee shop" → "...standing..."




"a boy, standing on the beach, t-shirt, sunset, full body" → "... hands in hands ..." + 



"A boy, standing, street, long pants" → "...running..."



"1girl, white medium hair, looking at viewer, jacket, outdoors, full body" → "... raising hands ..." + 



Experiments

Synthesis Results

Generated Image

Ours

Fixed Seed

P2P

SDEdit (0.5)

SDEdit (0.8)

PnP



"An apple on the table" → "Two apples ..."



"A kitten is sitting on the floor" → "... laying ..."

Real image editing results

Input Real Image

Ours

Fixed Seed

P2P

SDEdit (0.5)

SDEdit (0.8)

PnP



"A photo of a running corgi"



"A photo of a person, black t-shirt, raising hand"

Ablation Study

“A horse facing camera” → “...side view...”



synthesis with P_g



synthesis with P



step 0



step 5



step 15



step 30

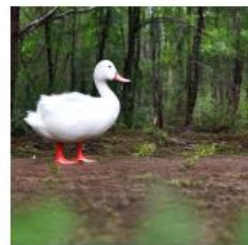


step 45

Denoising steps →

(a) Results of mutual self-attention control starting from different denoising steps

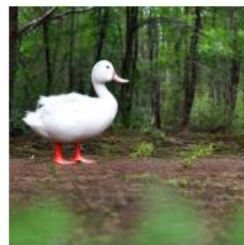
“a duck” → “... sitting ...”



synthesis with P_g



synthesis with P



layer 0~15

Whole U-Net



layer 0~3

Encoder



layer 4~7



layer 8~10

Decoder



layer 10~15

25

Results with T2I-Adapter

Generated Image



Ours

Fixed seed



"A bear is walking in forest" +  → "... standing ..." + 



"A photo of a dog, standing in Times Square, highly detailed" +  → "... sitting ..." + 

Input Real Image


Ours

Fixed seed



"A realistic photo of a sitting cat, camera view, masterpiece, best quality" + 



"A realistic photo of a horse, standing on its hind legs, grassland" + 

Extension to Video Synthesis

Generated Image



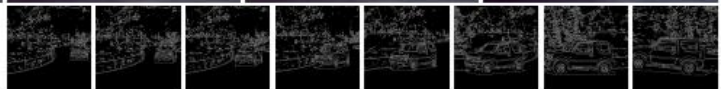
Results with MasaCtrl



“A bear dancing on the street, realistic photo, masterpiece, best quality” +



“A car is moving on the road, realistic photo, masterpiece, best quality” +

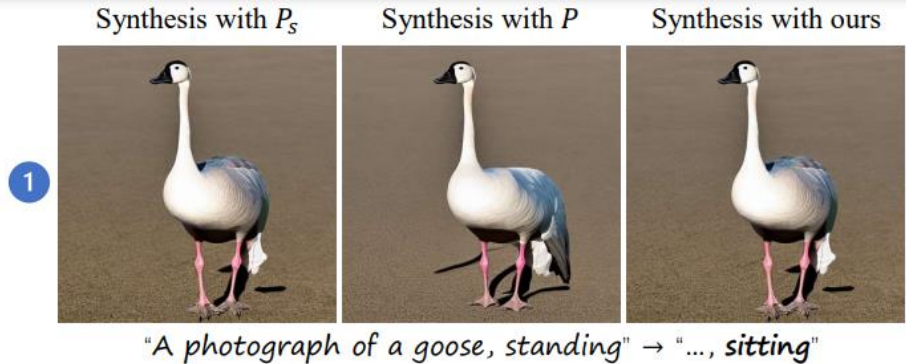




Conclusion

Limitations

1. Relies on the image layout synthesized from the target prompt P, it would fail if the SD model could not generate a desired layout or shape.
2. This method will fail when the target image contains unseen content or the target image layout/structure changes drastically.
3. There still are some slight differences between the source image and the edited image.





**Thanks
for Watching!**