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

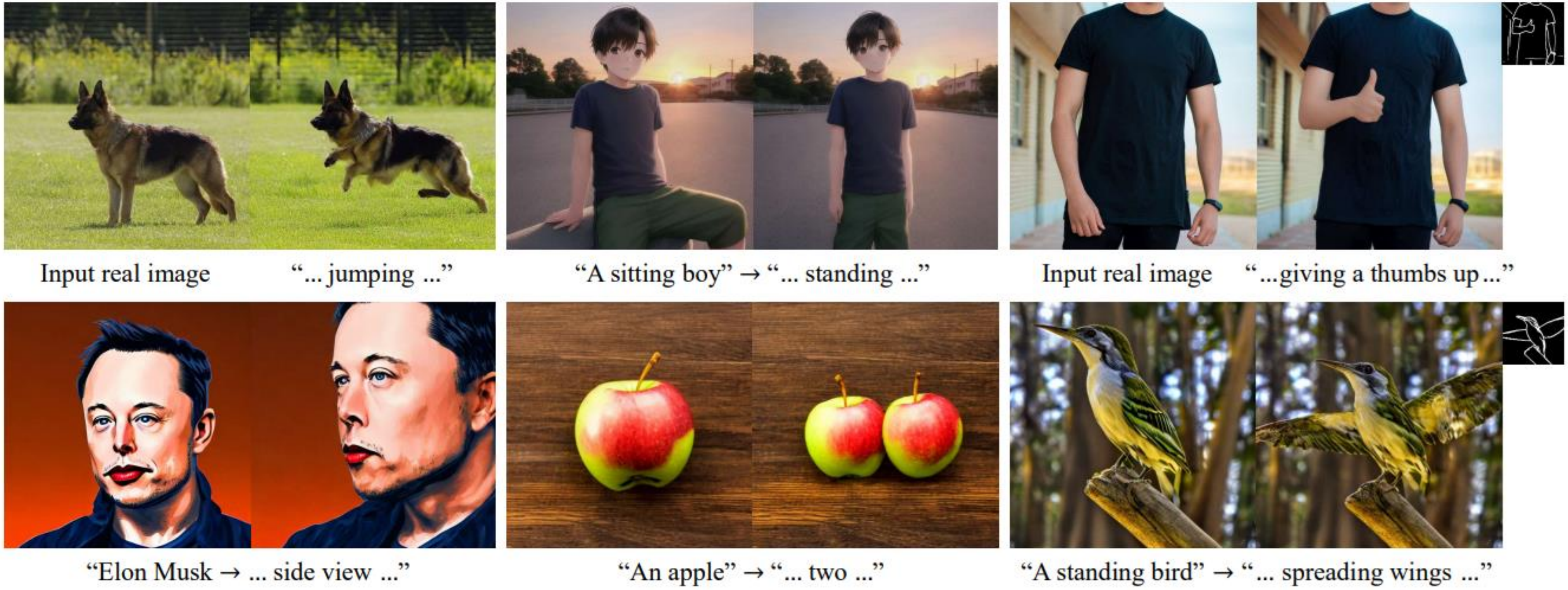
Mingdeng Cao[1,2*] Xintao Wang[2✉] Zhongang Qi[2] Ying Shan[2] Xiaohu Qie[2] Yinqiang Zheng[1✉]

[1]The University of Tokyo [2]ARC Lab, Tencent PCG

ICCV 2023

Introduction

Abstract

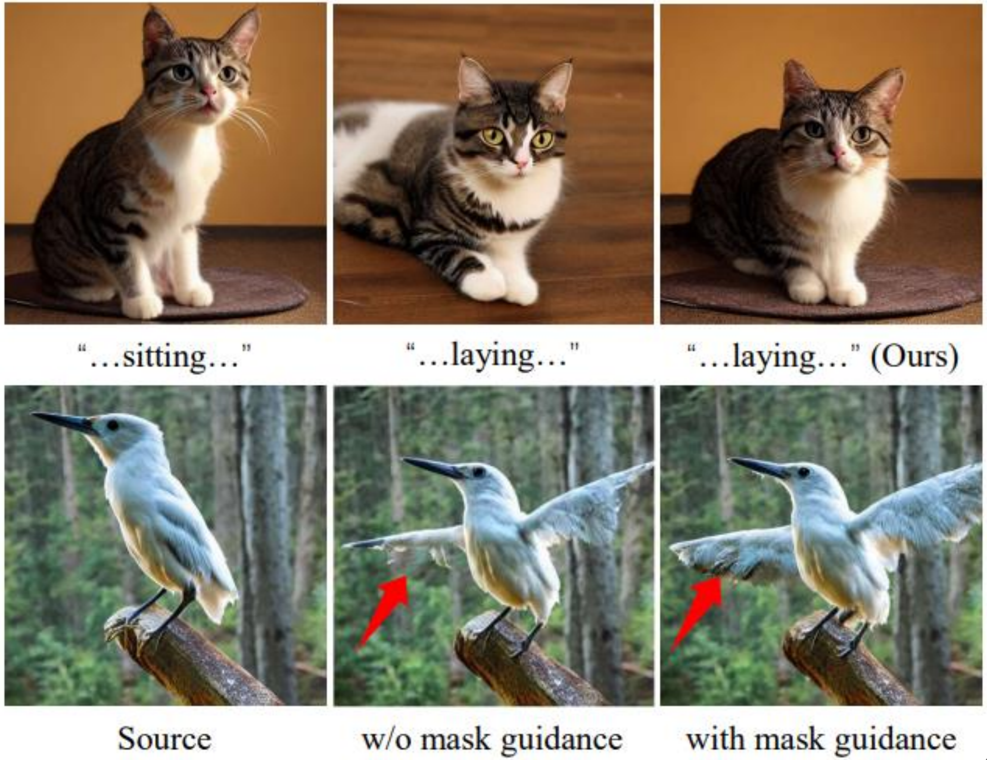


MasaCtrl can perform text-based non-rigid image synthesis and real image editing without finetuning.

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**.



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

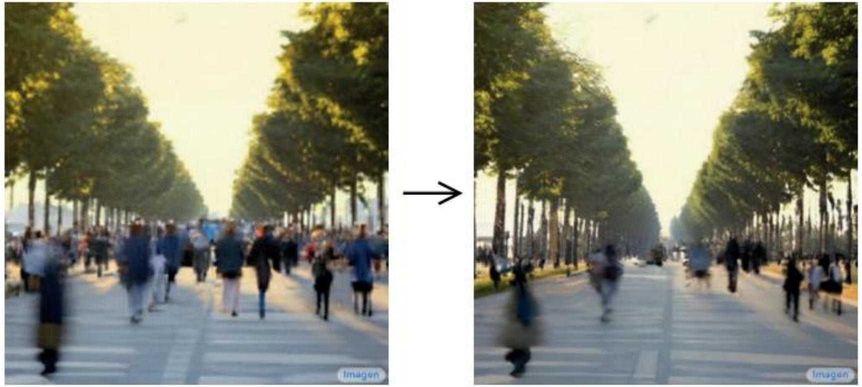
DiffusionCLIP

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

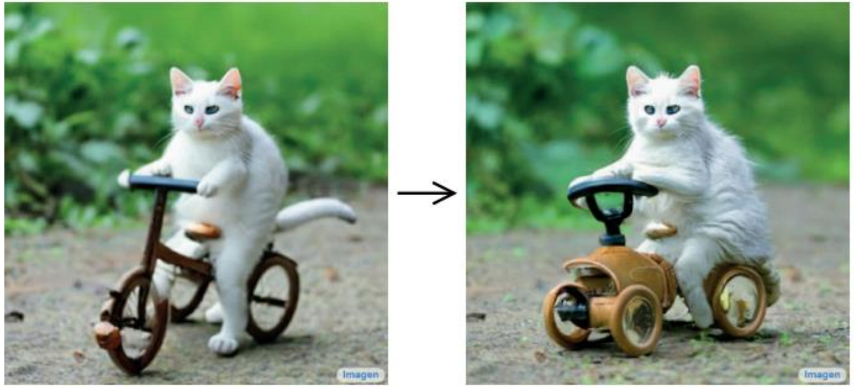
Text-guided Image Editing

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.**



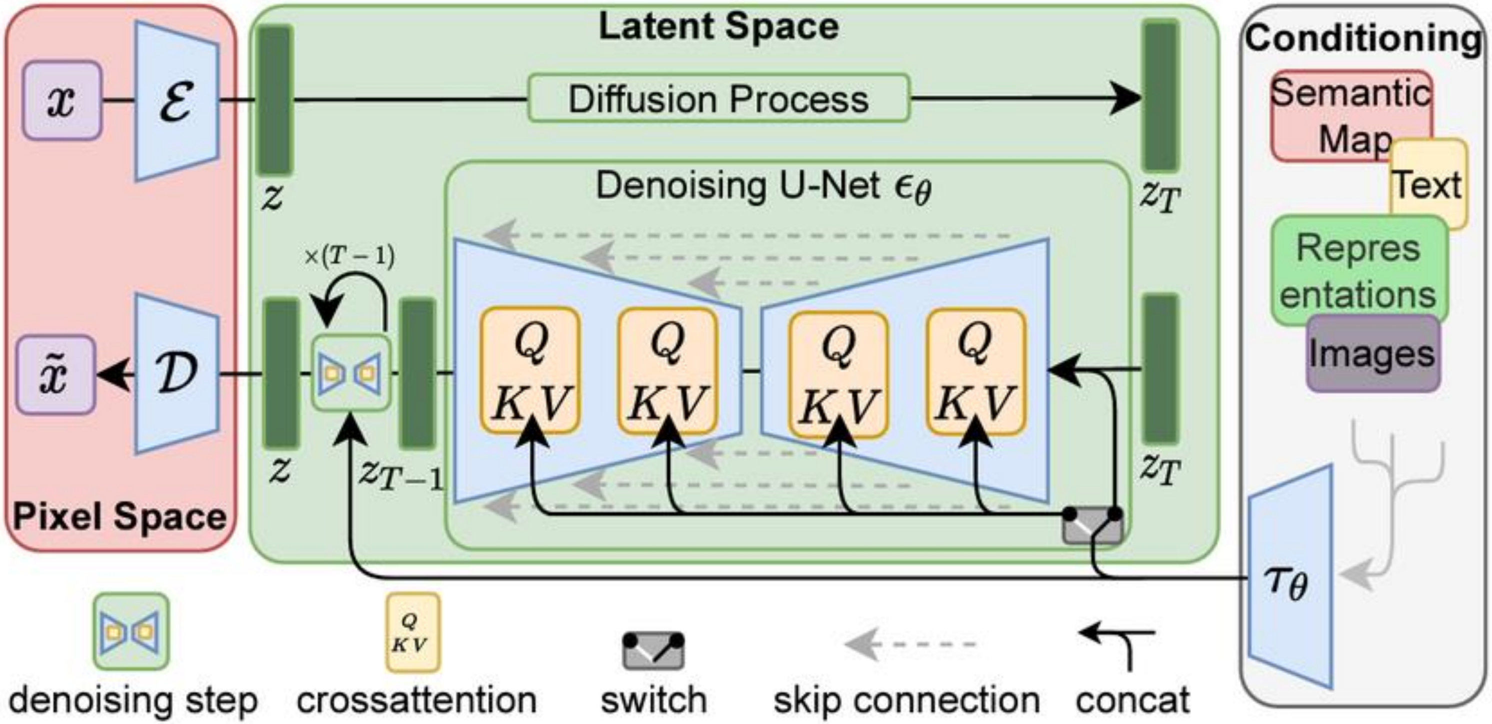
"The boulevards are crowded today."



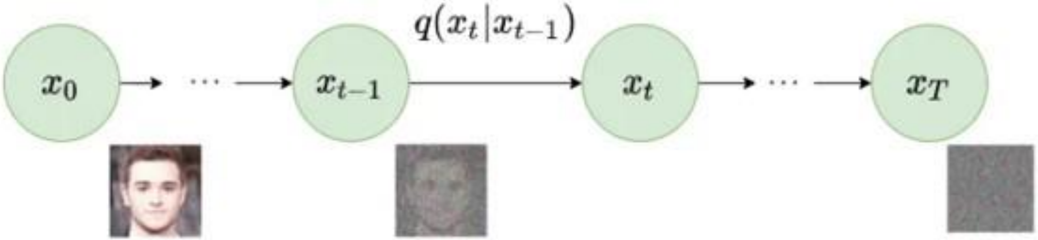
"Photo of a cat riding on a ~~bicycle~~ car."

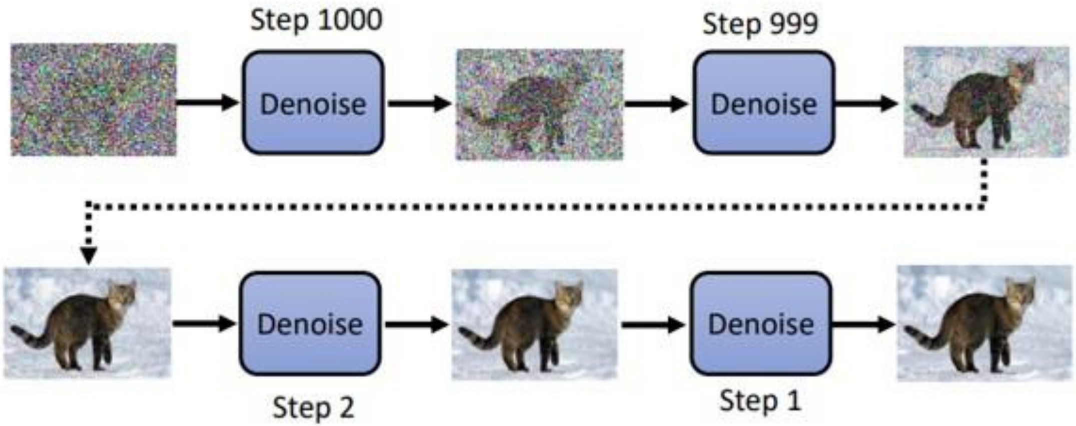
Preliminaries

Stable Diffusion

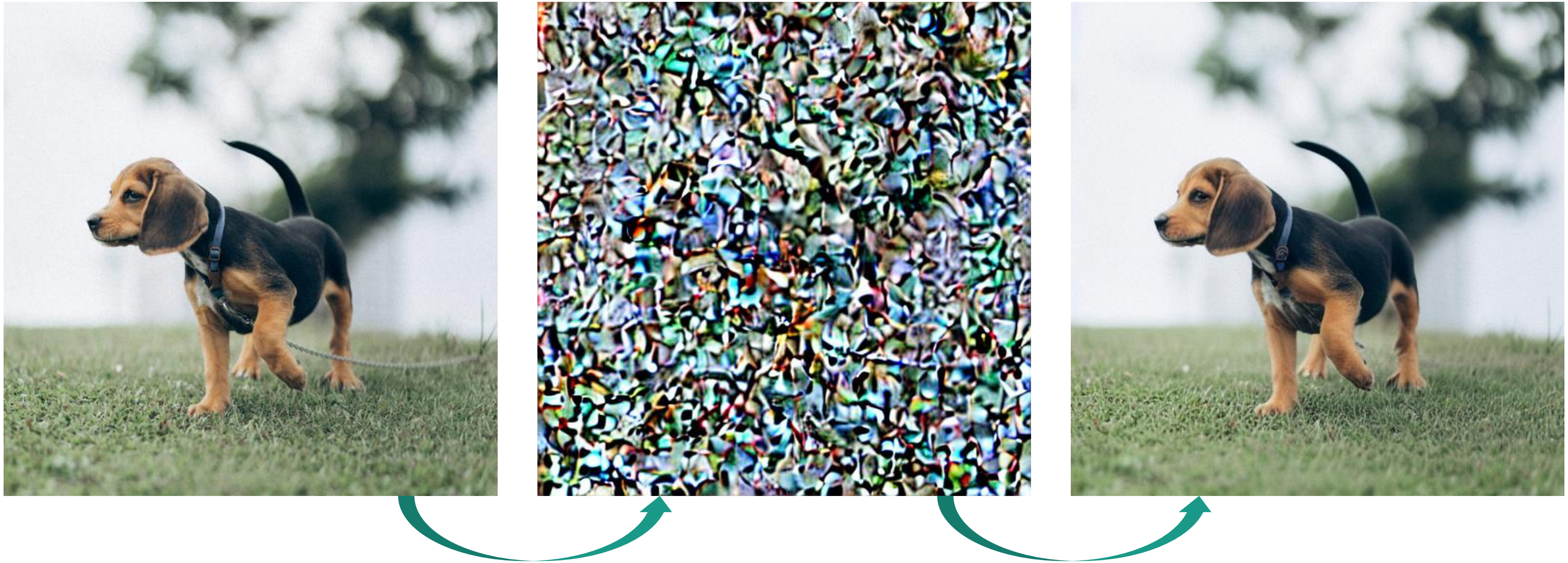


Diffusion Model

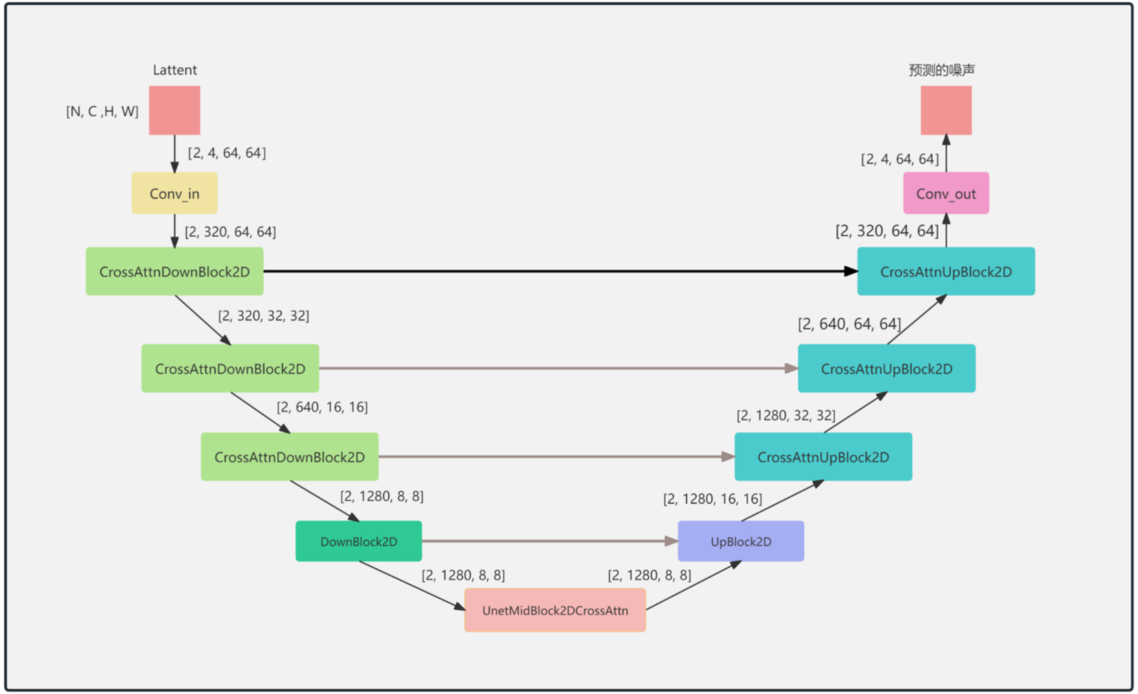


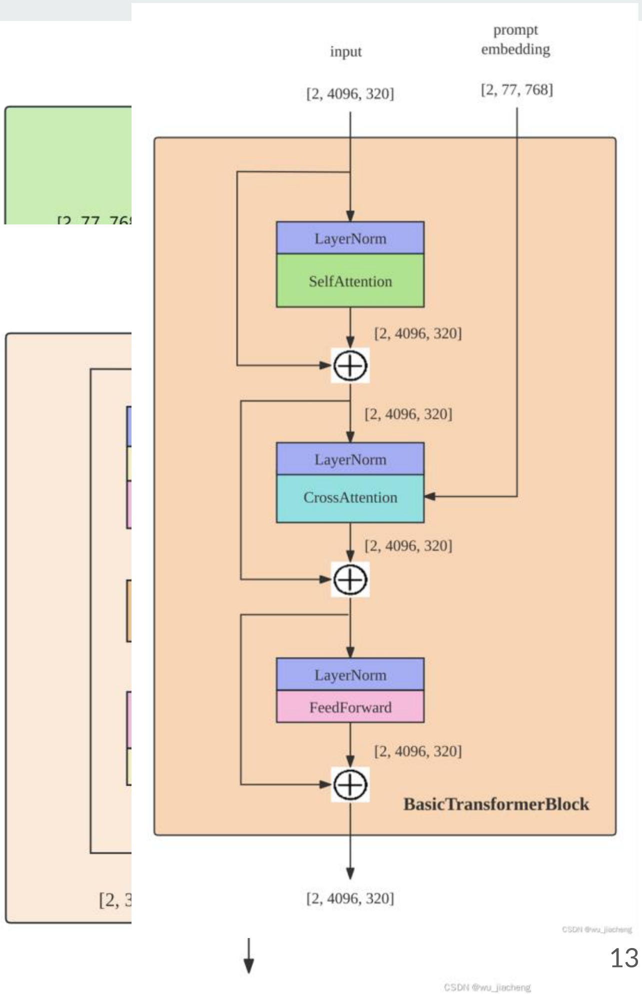


DDIM Inversion



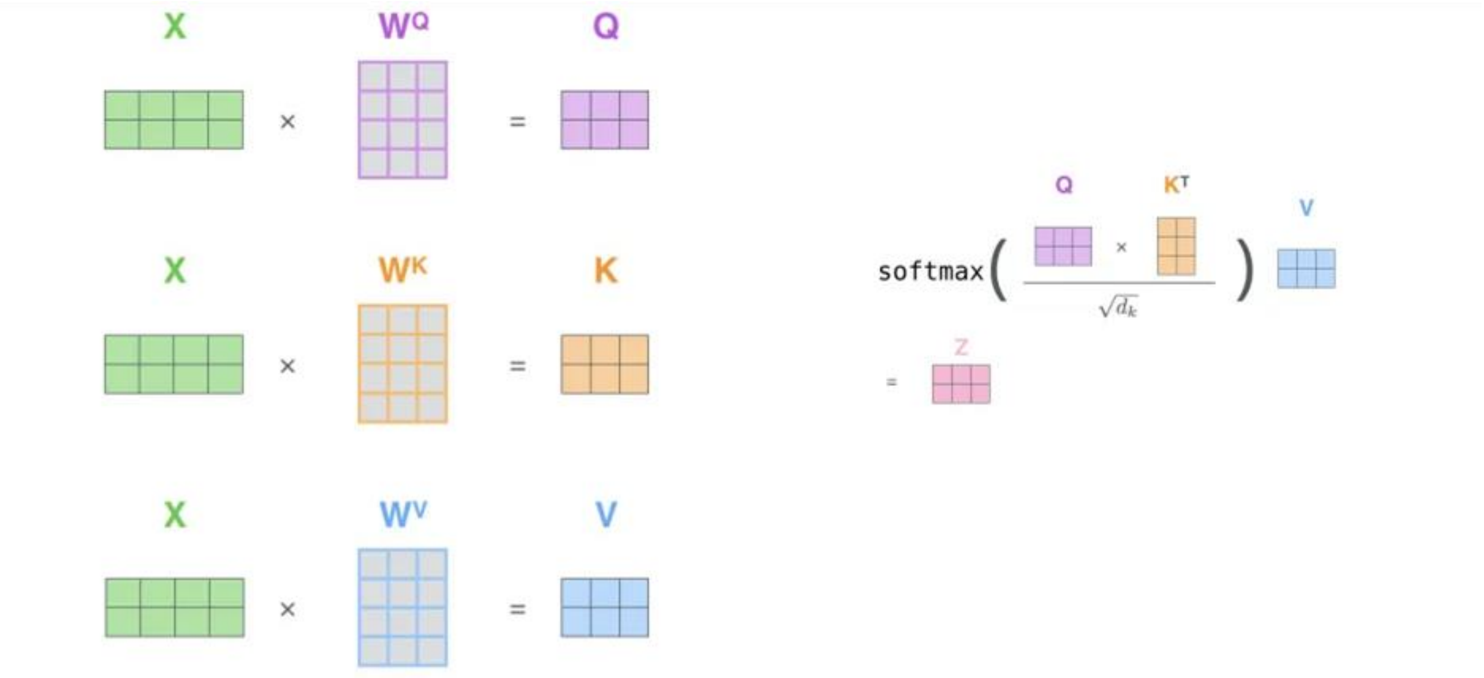
Attention Block



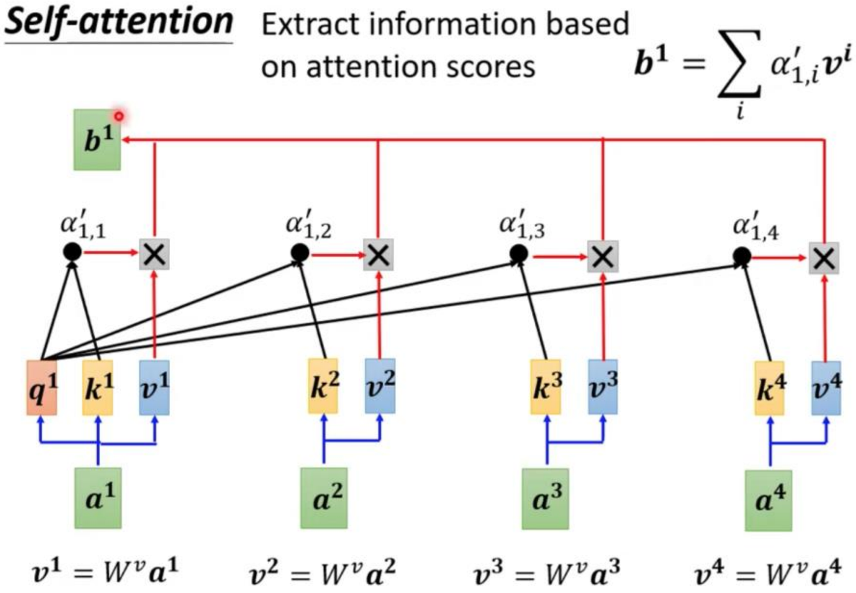


Attention Mechanism

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

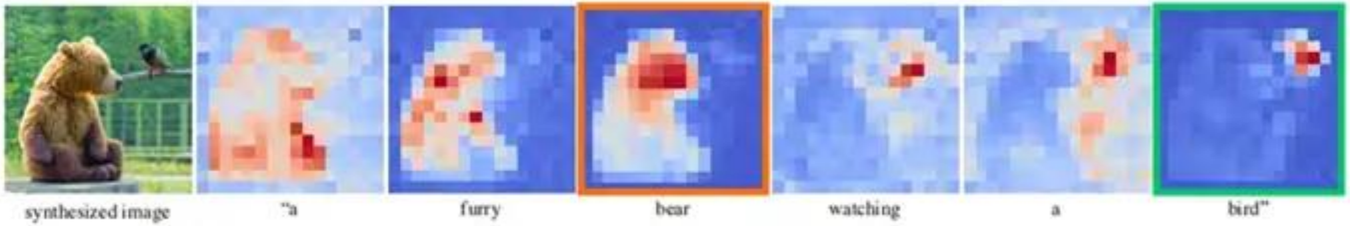


Attention in Stable Diffusion



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^l_t** .

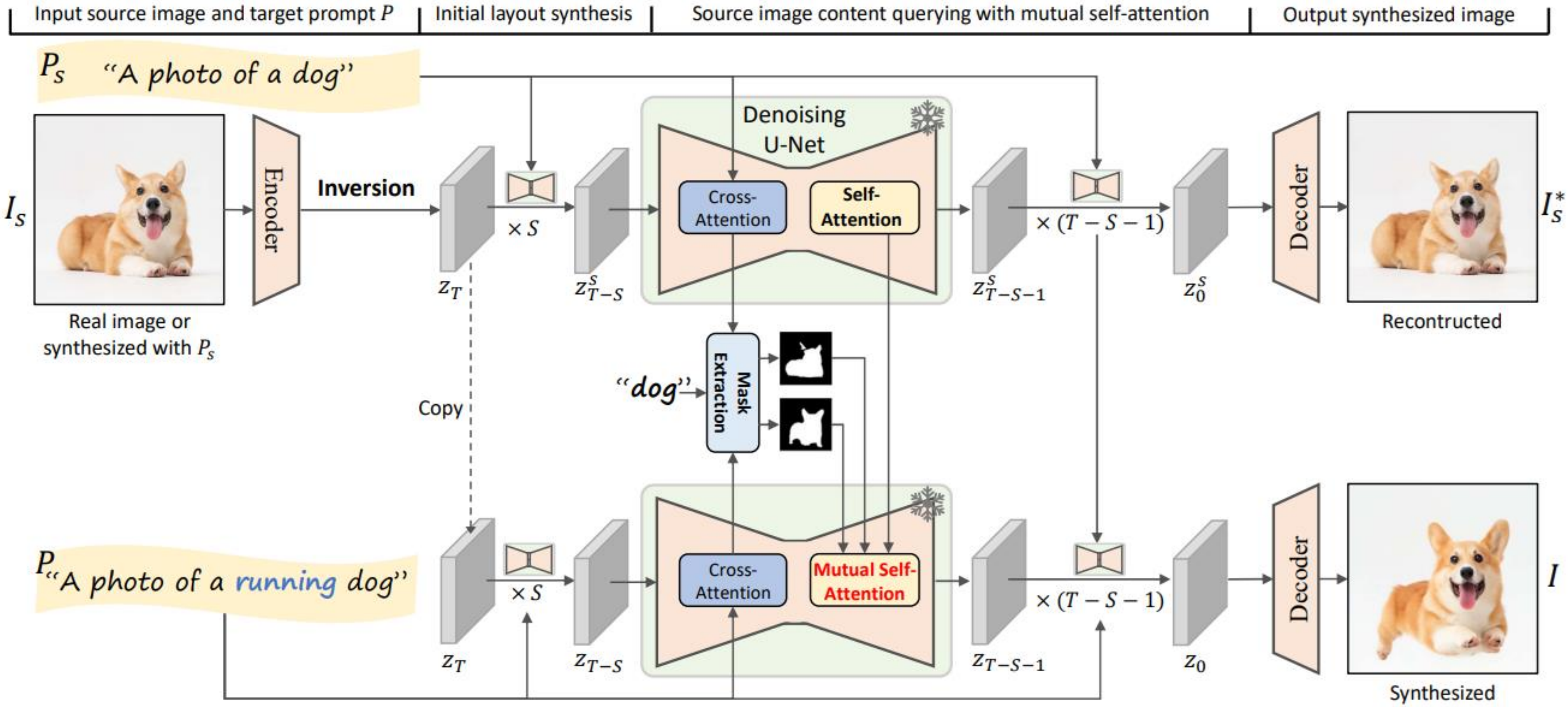
- Then they are reorganized by a **self-attention** layer.
- Receive textual information from the given text prompt P by the following **cross-attention** layer.



Average attention maps across all timestamps

Method

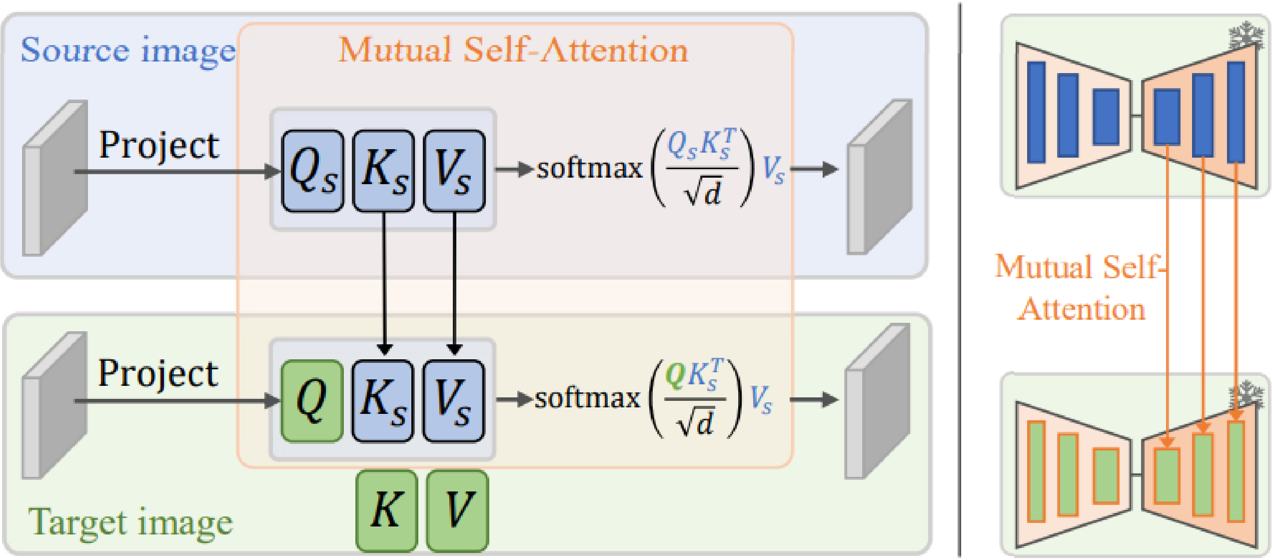
Pipeline



Input source image and target prompt P
Initial layout synthesis
Source image content querying with mutual self-attention
Output synthesized image
P_s "A photo of a dog"
I_S
Real image or synthesized with P_s
Encoder
Inversion
z_T
$\times S$
z^S_{T-S}
Denoising U-Net
Cross-Attention
Self-Attention
z^S_{T-S-1}
$\times (T-S-1)$
z^S_0
Decoder
I^*_S
Recontructed
Copy
"dog"
Mask Extraction
P "A photo of a running dog"
z_T
$\times S$
z_{T-S}
Cross-Attention
Mutual Self-Attention
z_{T-S-1}
$\times (T-S-1)$
z_0
Decoder
I
Synthesized

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**.

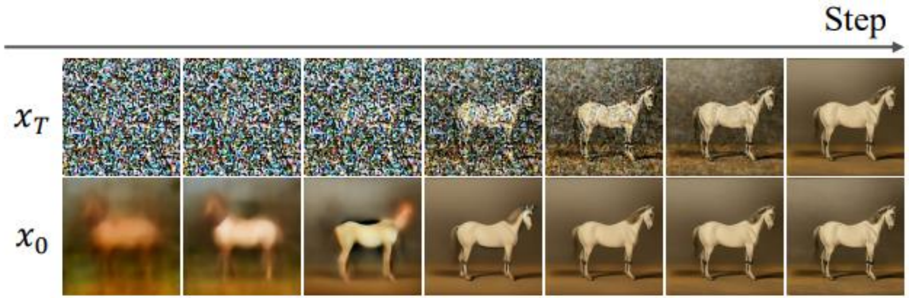


1. Mutual Self-Attention

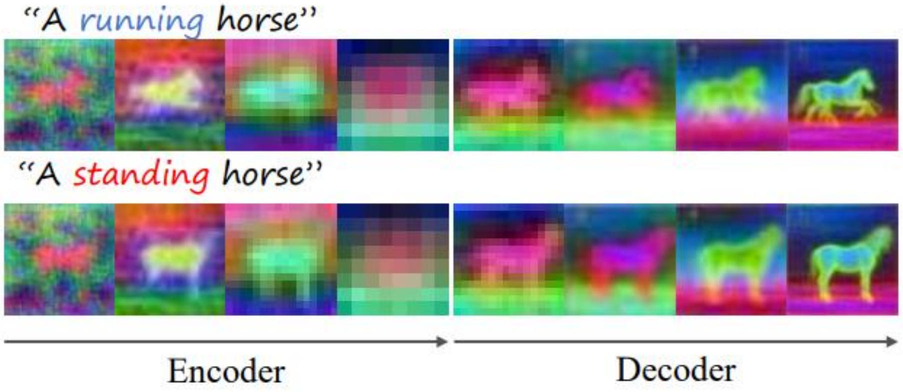
However, intuitively performing such attention control on **all layers** among **all denoising steps** will result in an image I that is nearly the same as the reconstructed image Is.

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

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



(a) Intermediate results in denoising process



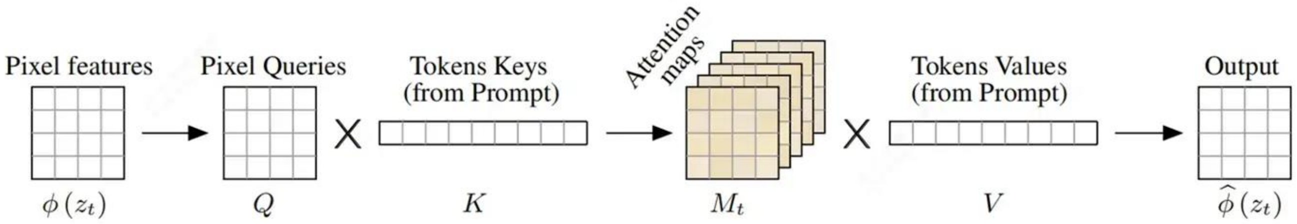
(b) *Query* feature visualization

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



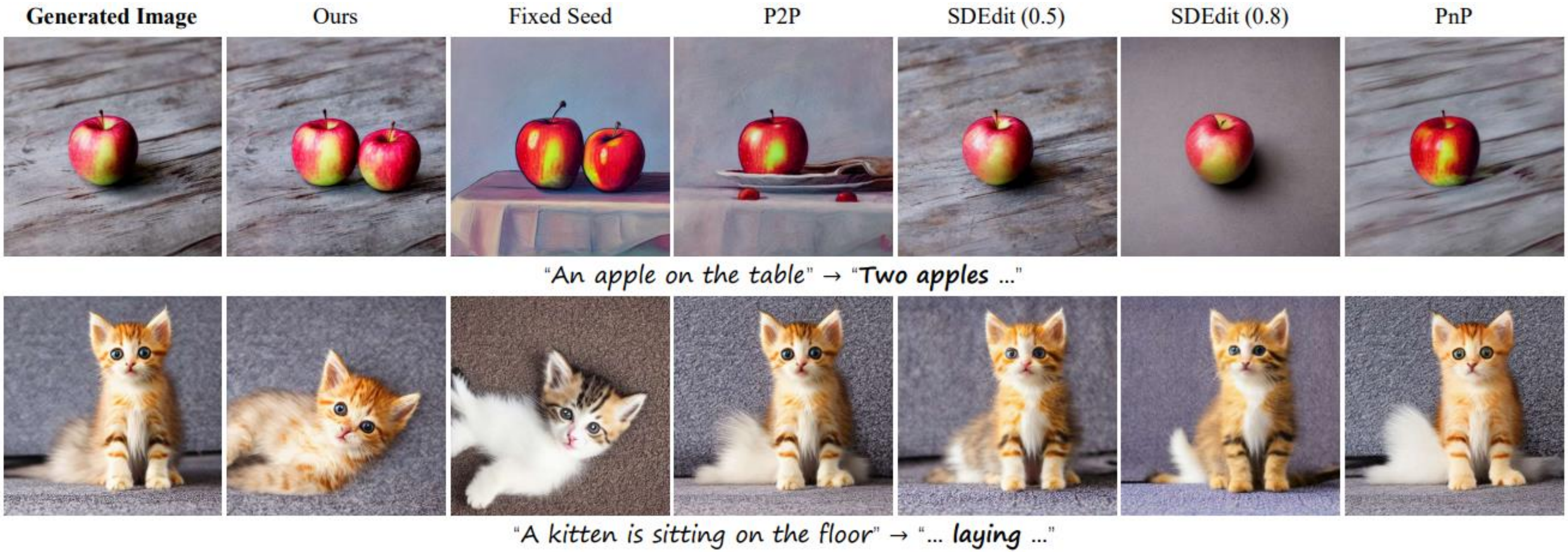
Text to Image Cross Attention

3. Integration to Controllable Diffusion Models

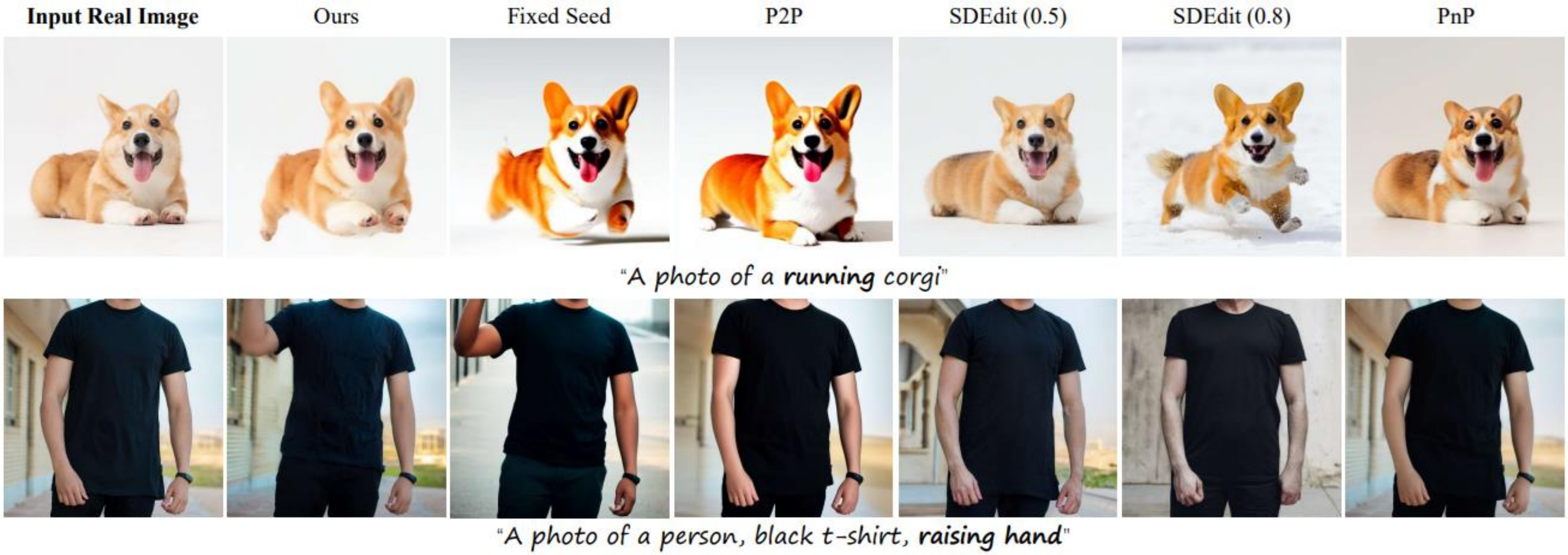
Our method can be easily integrated into existing controllable image synthesis method.

Experiments

Synthesis Results

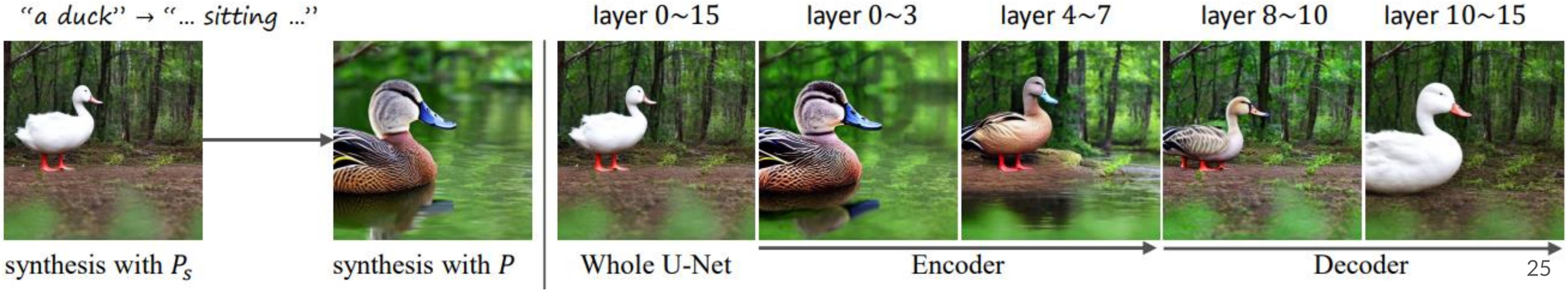


Real image editing results

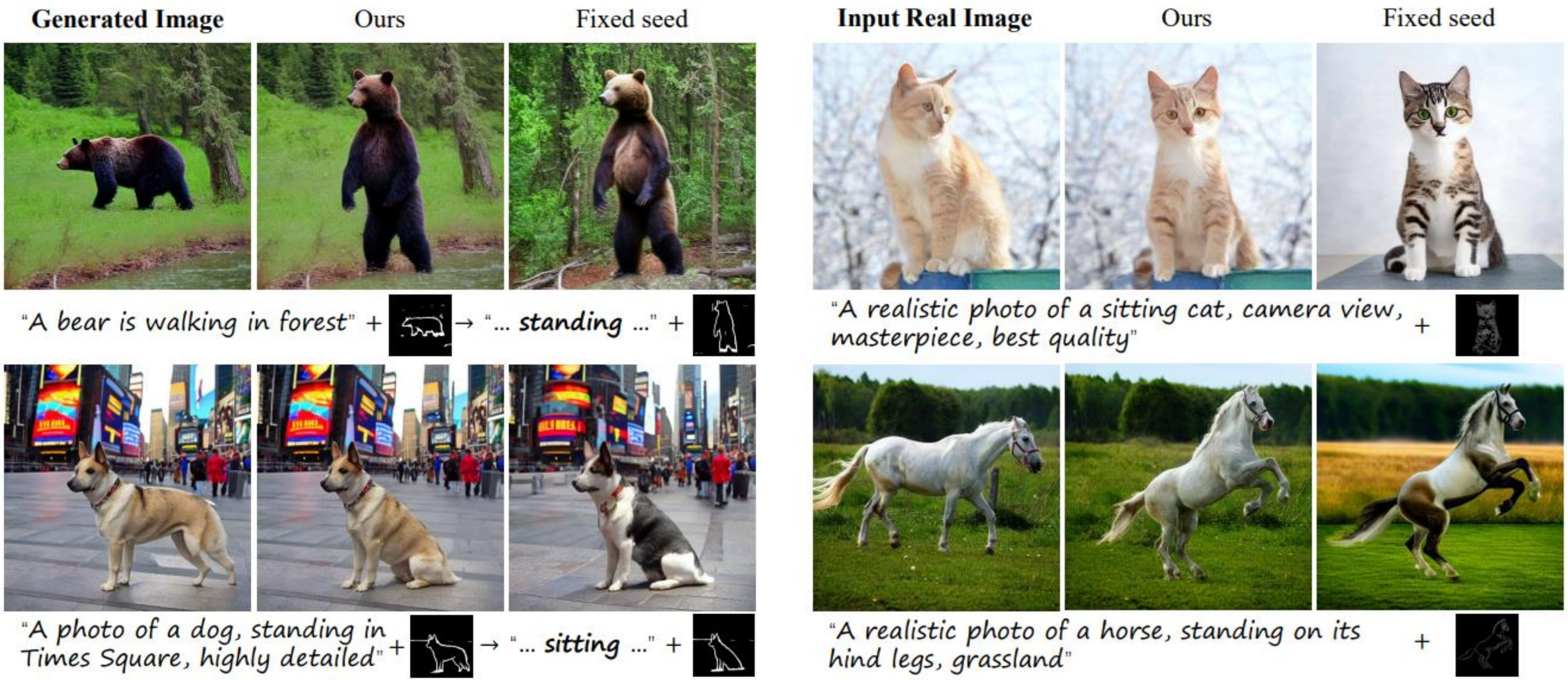


Ablation Study

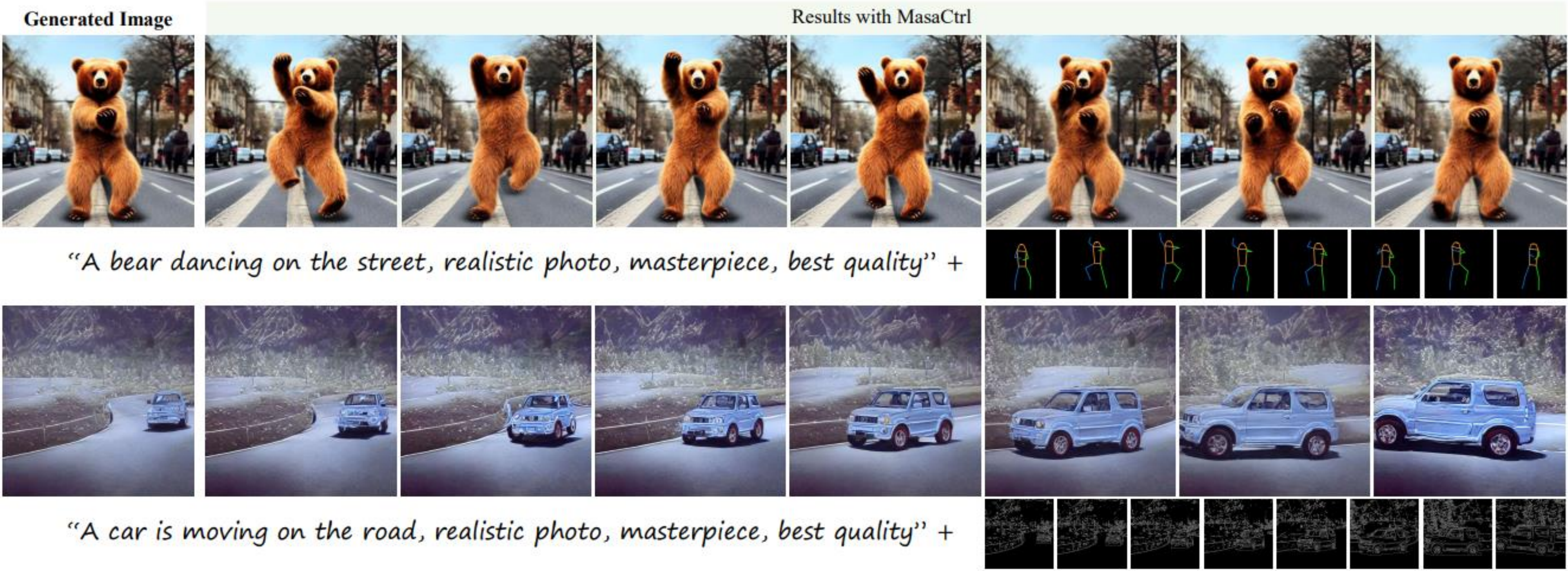
(a) Results of mutual self-attention control starting from different denoising steps



Results with T2I-Adapter



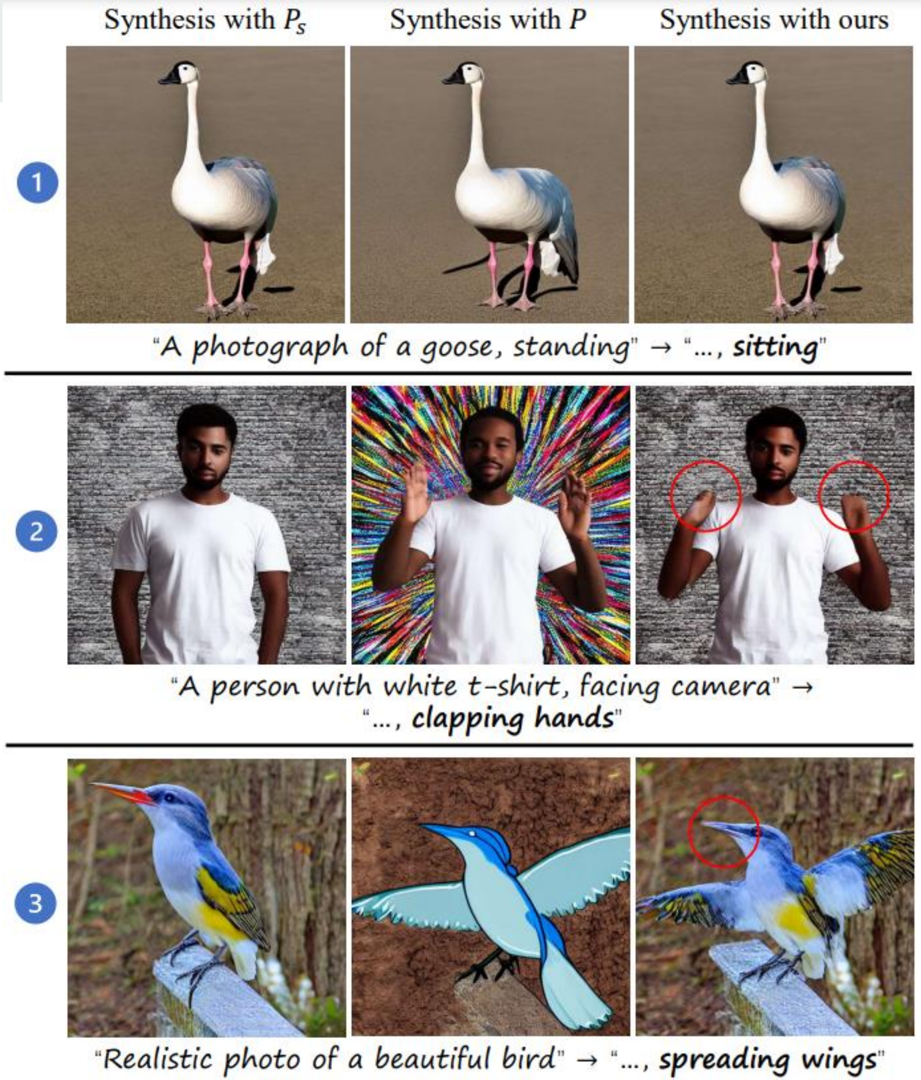
Extension to Video Synthesis



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!