Diffusion features for image editing and beyond

24/10/28

Junyi Zhang

Content

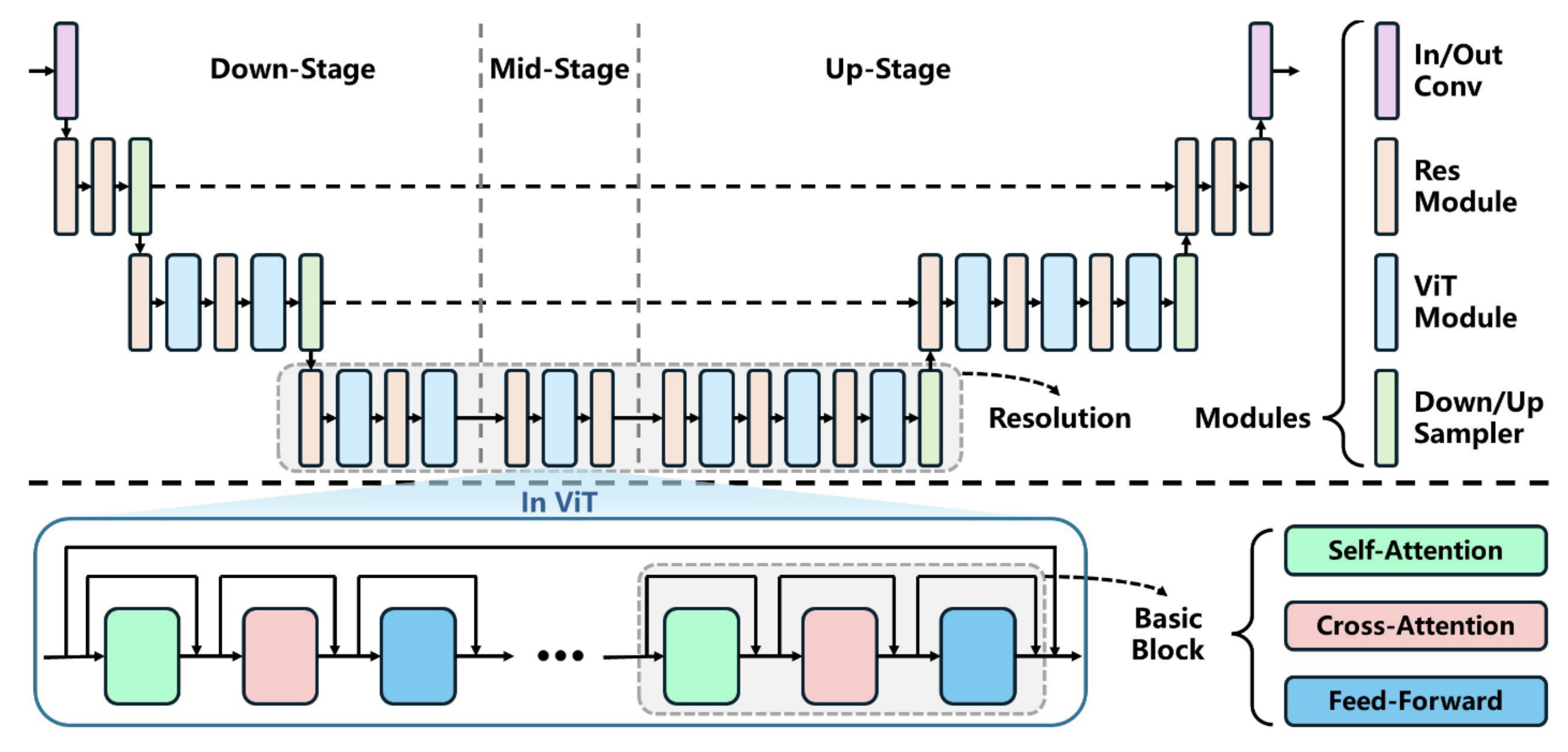
- Cross-attention features for image editing
  - Prompt-to-Prompt (Neurips22)

- Residual / Self-attention features for image editing
  - PnP-Diffusion (CVPR23)

- Diffusion features for perception tasks

Features of stable diffusion models

- U-Net -> Down/Mid/Up Block -> Res/ViT layer -> Self/Cross-attention



Prompt-to-Prompt Image Editing with Cross Attention Control

Task overview

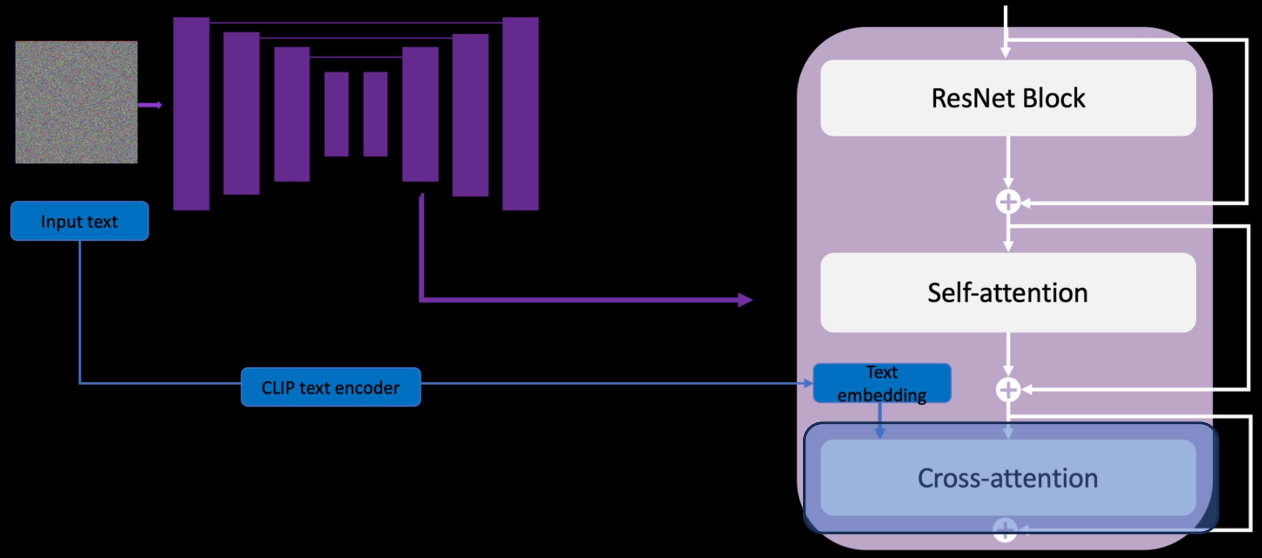
- Generate an image with the prompt
- -> edit the generated image by updating the prompt

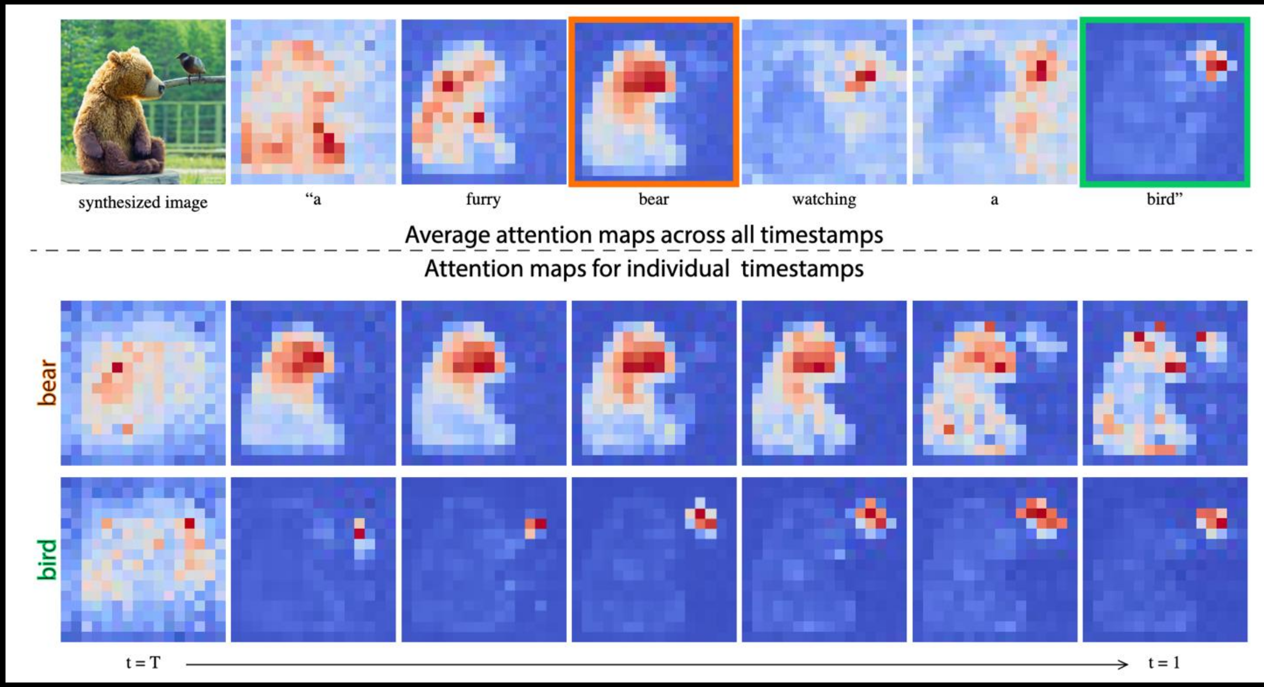
Motivation

- Editing the generated image by "using the same random seed"
- -> structure completely changed

Motivation

- The cross-attn maps of a text-conditioned diffusion model connects the given prompt and generated image spatially





Method

- The cross-attention output is a weighted average of values V and the weights are the attention-maps M: $\emptyset(z_t) = MV$
- One can manipulate the attention map to edit the generated image

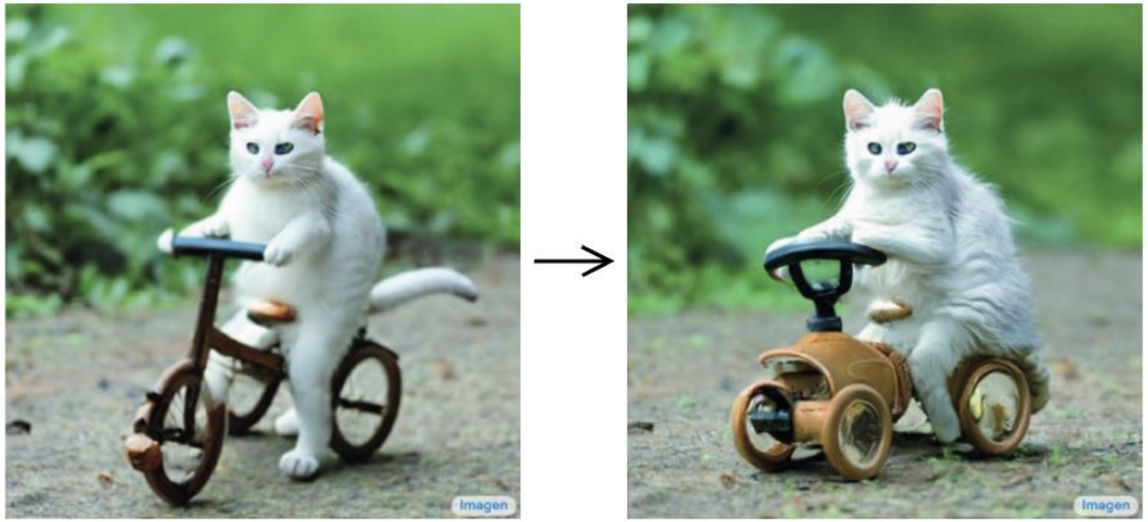
Algorithm 1: Prompt-to-Prompt image editing

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^*)

Method

- By choosing different ways of editing the attention maps, one can achieve various editing effects
- Word Swap

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

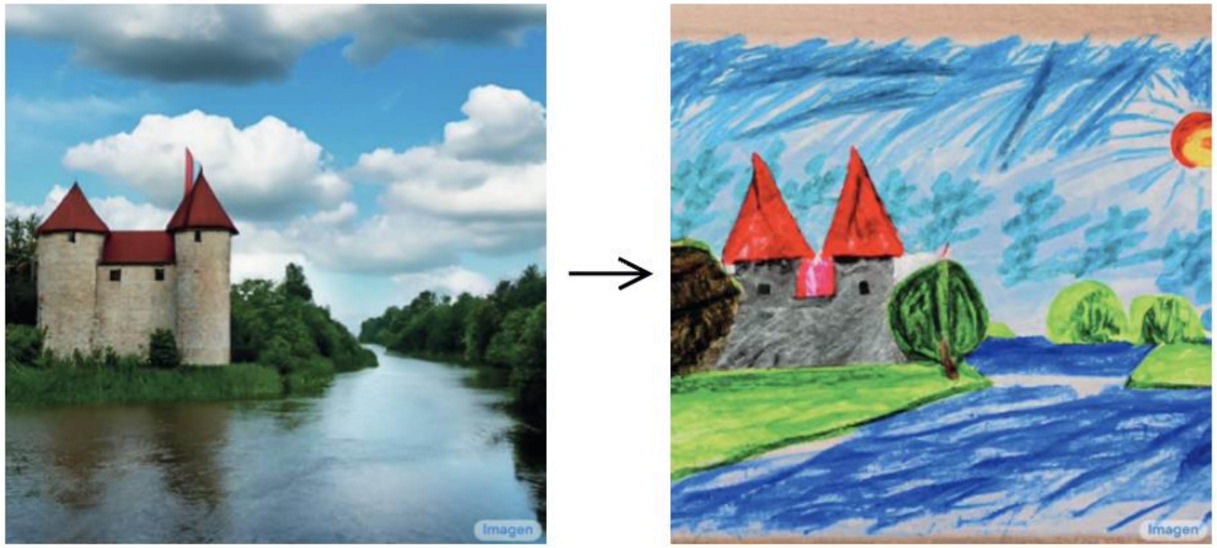


“Photo of a cat riding on a ~~bicycle~~.”
car

Method

- By choosing different ways of editing the attention maps, one can achieve various editing effects
- Adding a New Phrase

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = None \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

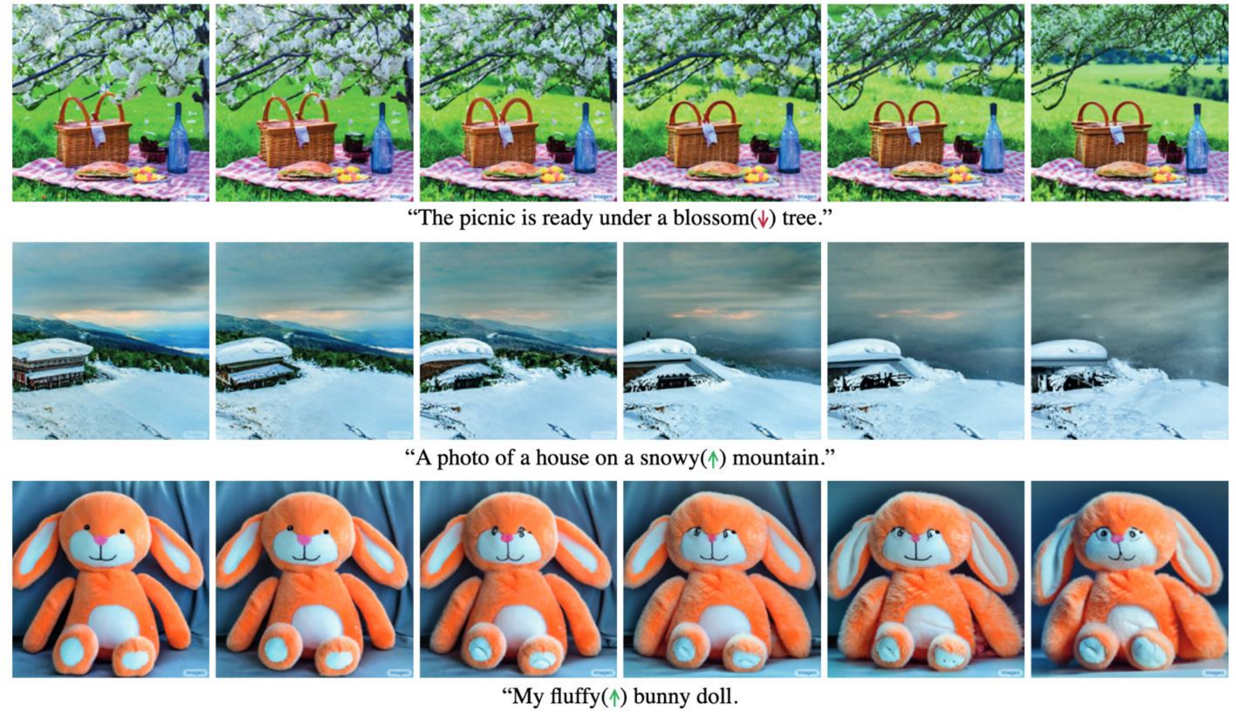


"Children drawing of a castle next to a river."

Method

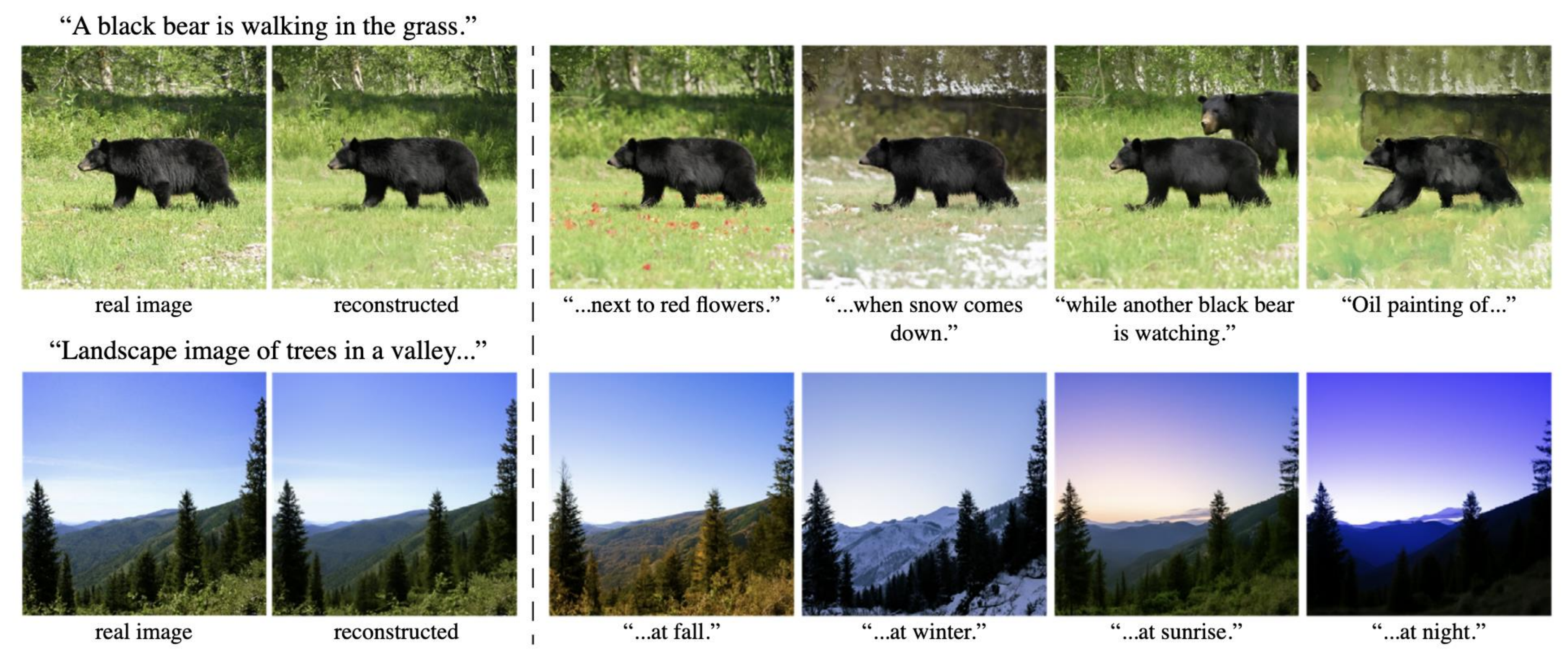
- By choosing different ways of editing the attention maps, one can achieve various editing effects
- Emphasizing / weakening certain words

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$



Editing on real images

- Only need to first invert the input image to latent space (z_T)

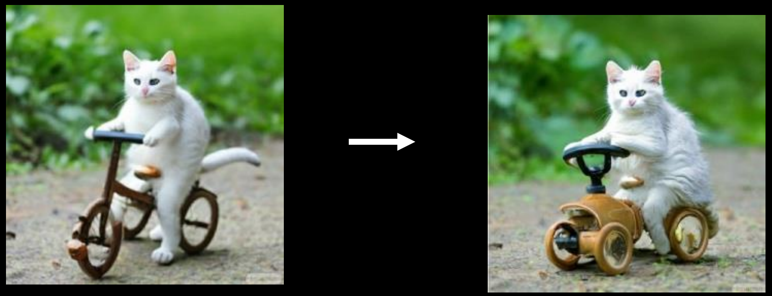


Limitation

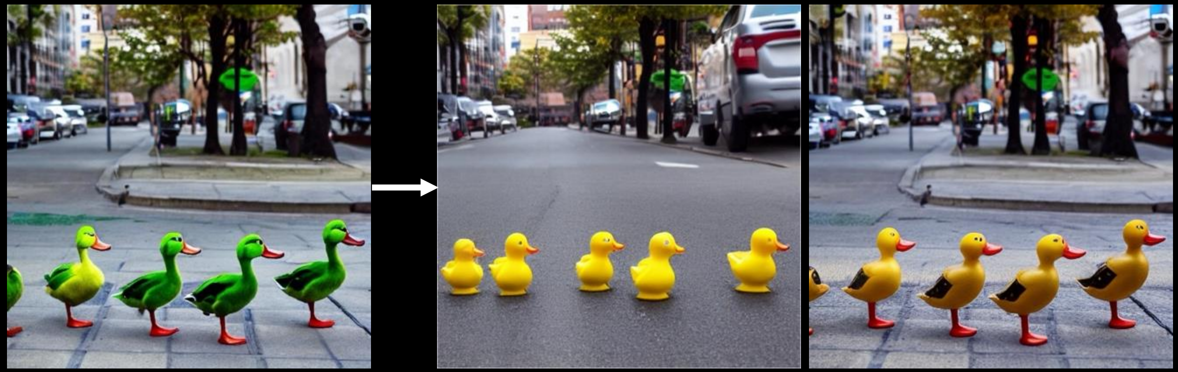
- <u>Prompt-to-Prompt</u>

Cross-attention manipulation

  - Limited in structure preservation
  - Limited to aligned source-target prompts



- <u>Pnp-diffusion</u>
  - fine-grained control over shape and layout
  - arbitrary source-target prompts



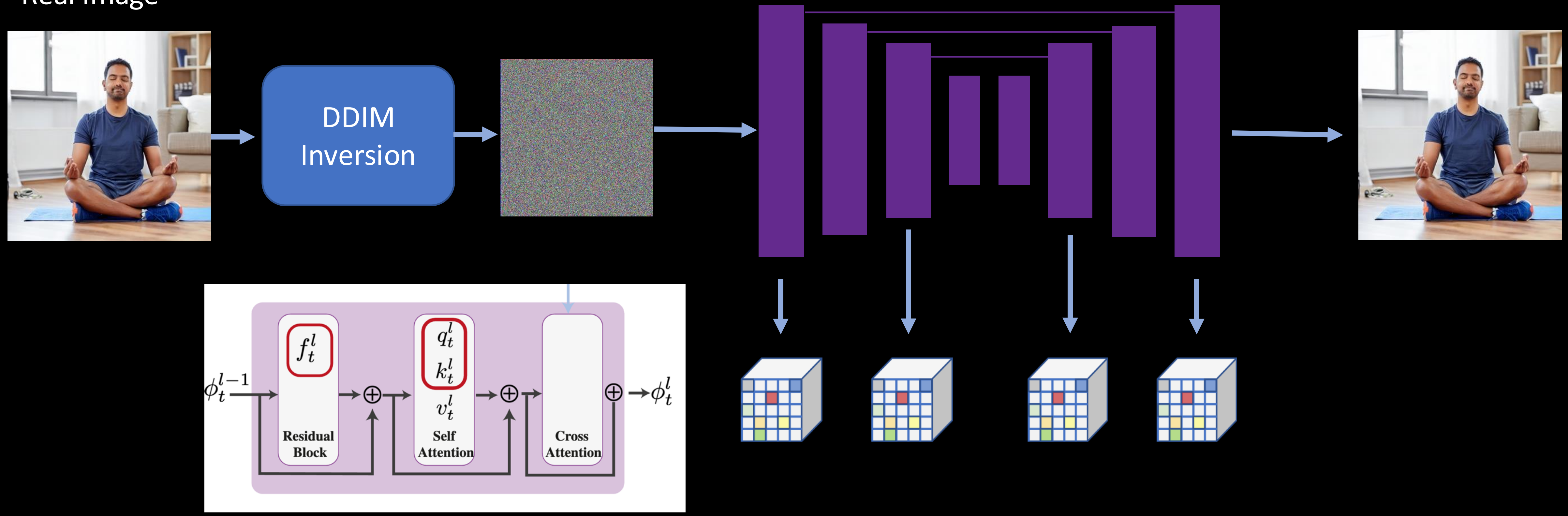
Prompt-to-Prompt Image Editing with Cross-Attention Control, ICLR 2023

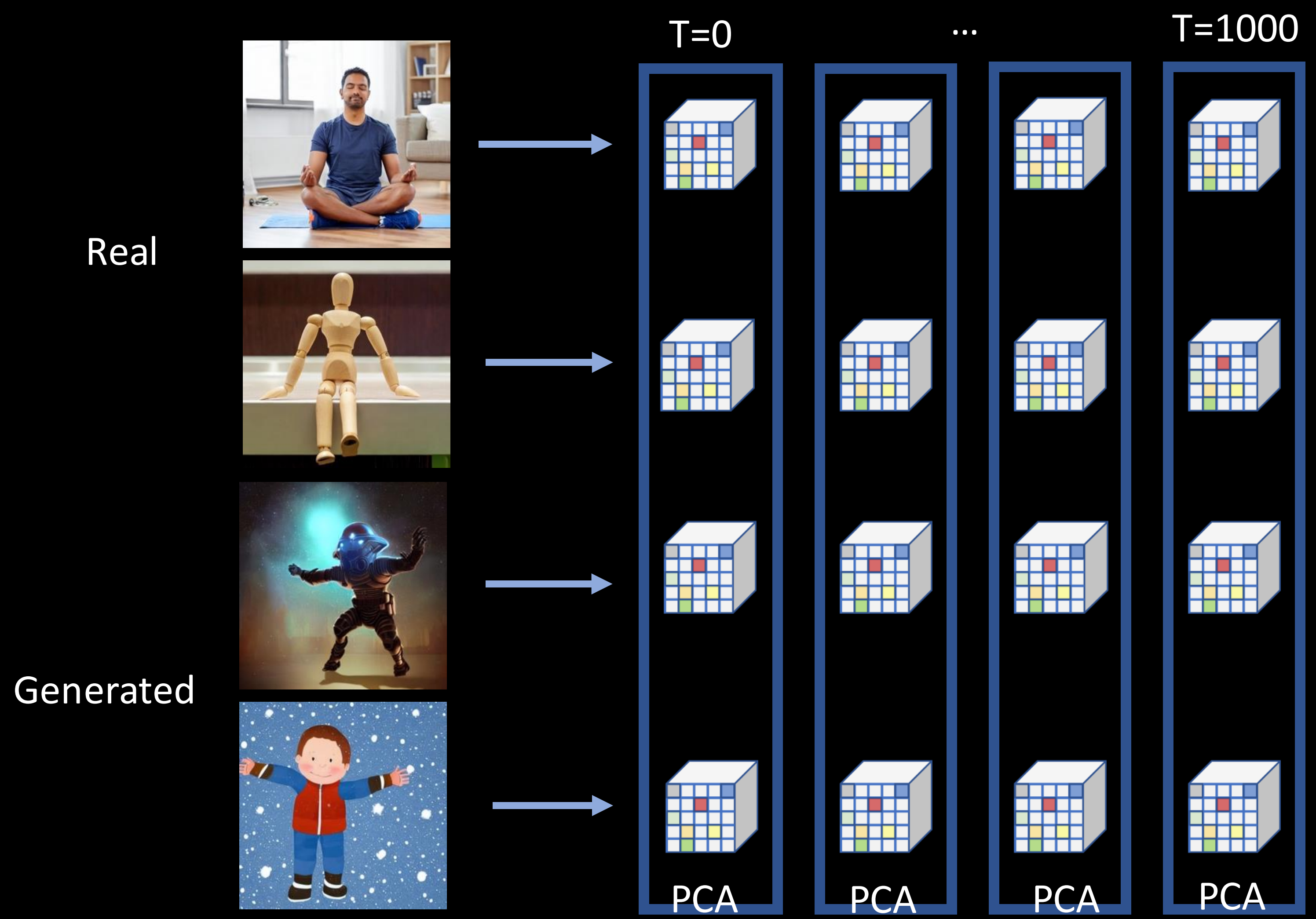
Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

1. How semantic layout is internally encoded in diffusion models?

2. How can we control structure in the generation process?

Real image


DDIM Inversion
f_t^l
q_t^l
k_t^l
v_t^l
ϕ_t^{l-1}
ϕ_t^l
Residual Block
Self Attention
Cross Attention



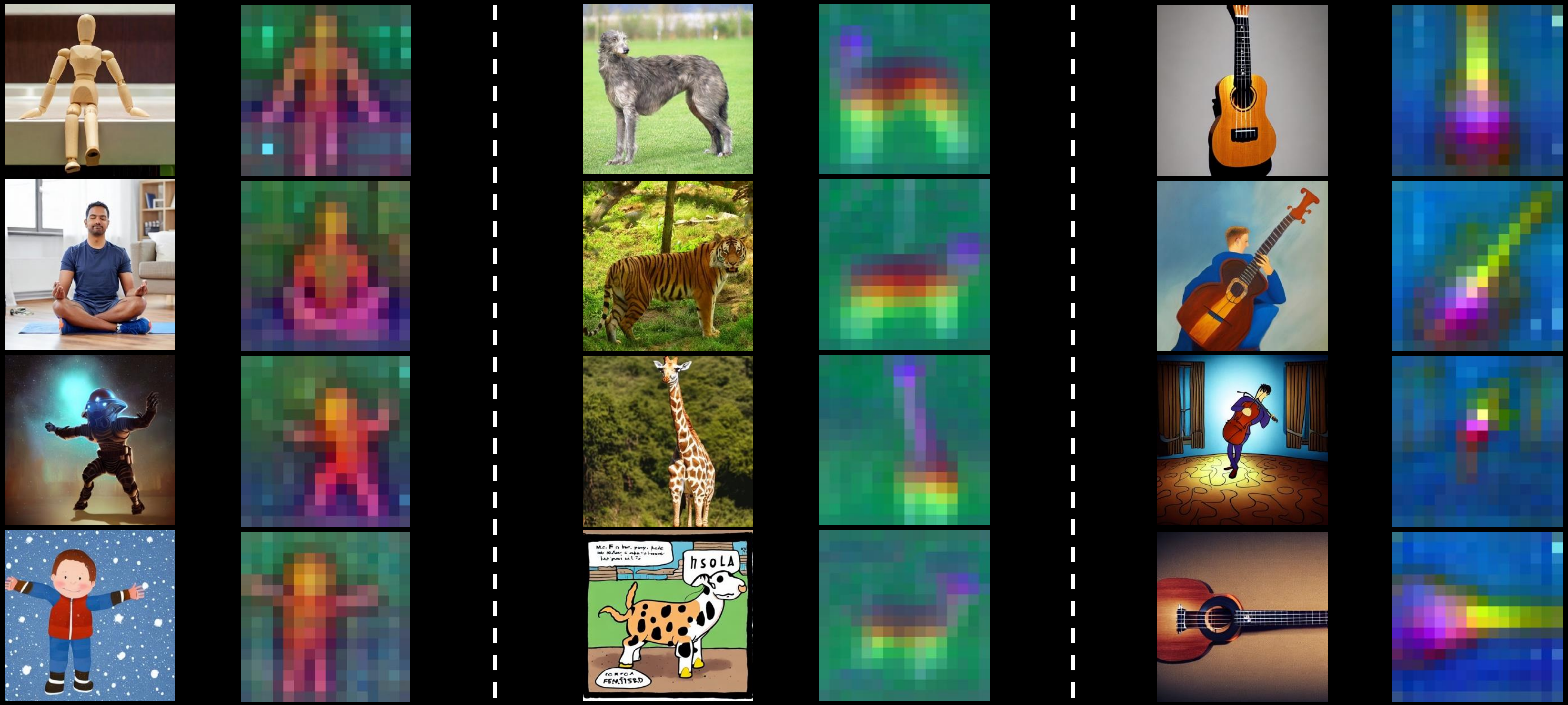
T=0
...
T=1000
Real
Generated
PCA
PCA
PCA
PCA

Top 3 PCA Componenets of ResBlock Features (Decoder Layer 4)

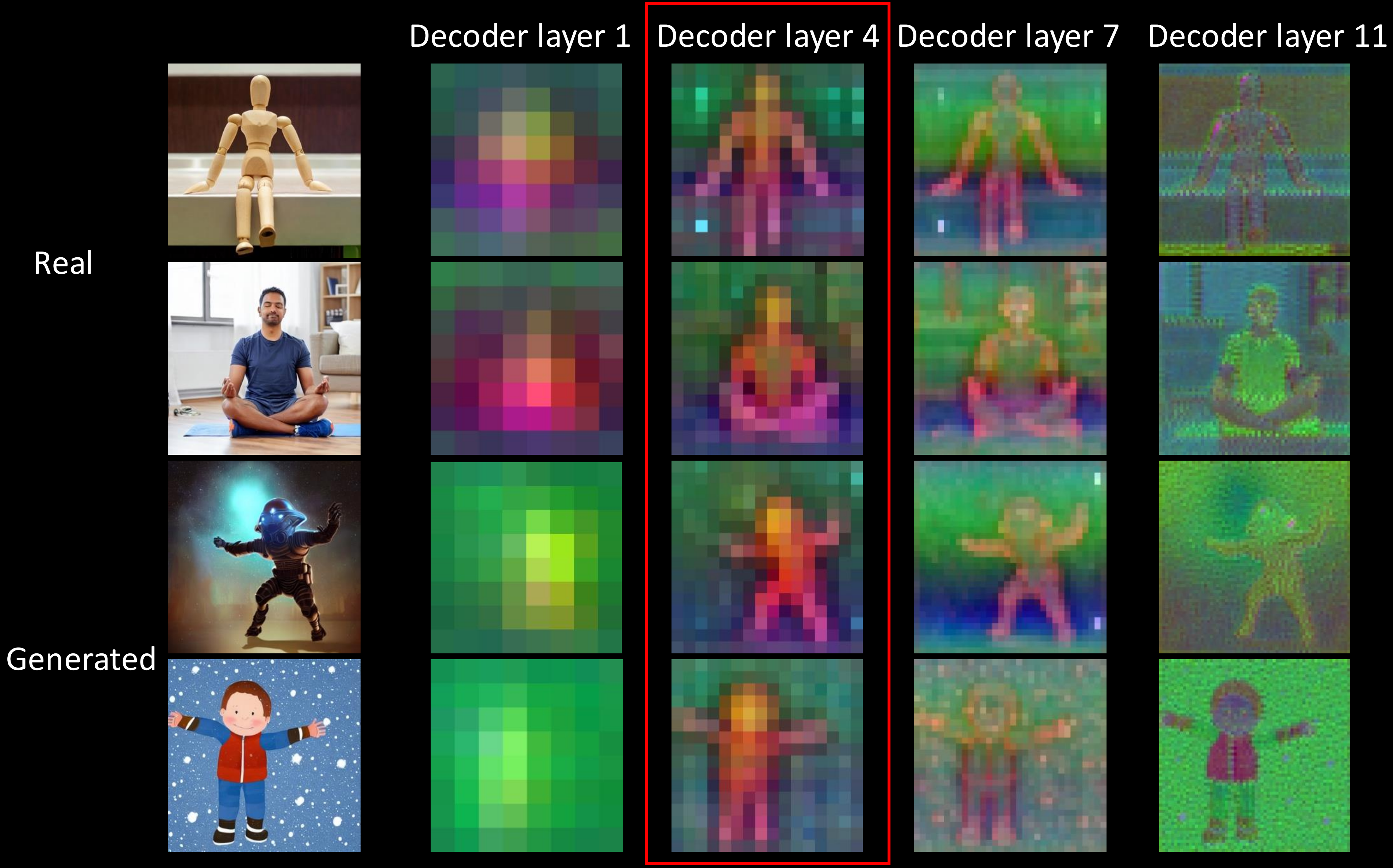
Humanoid images

Animal images

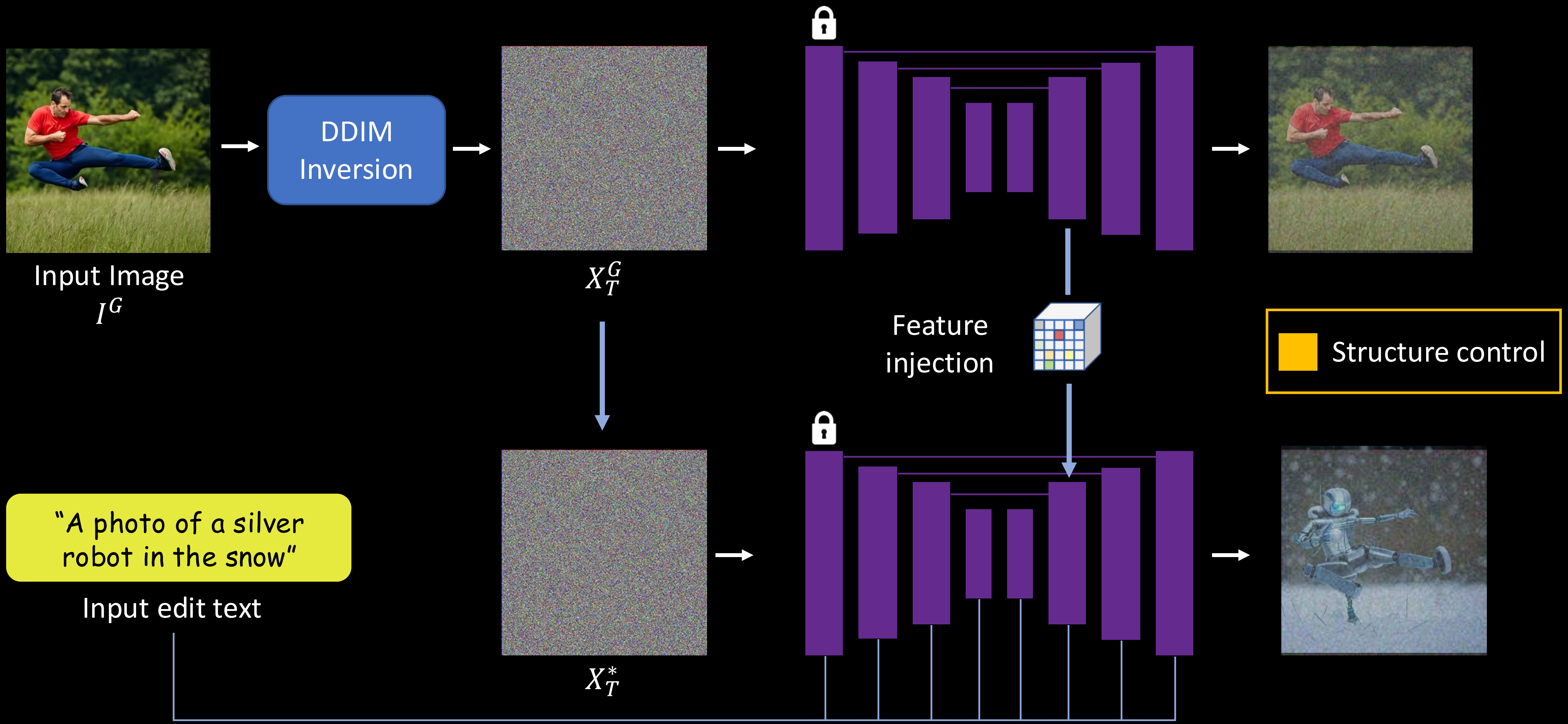
Instrument images



Top 3 PCA Componenets of ResBlock Features



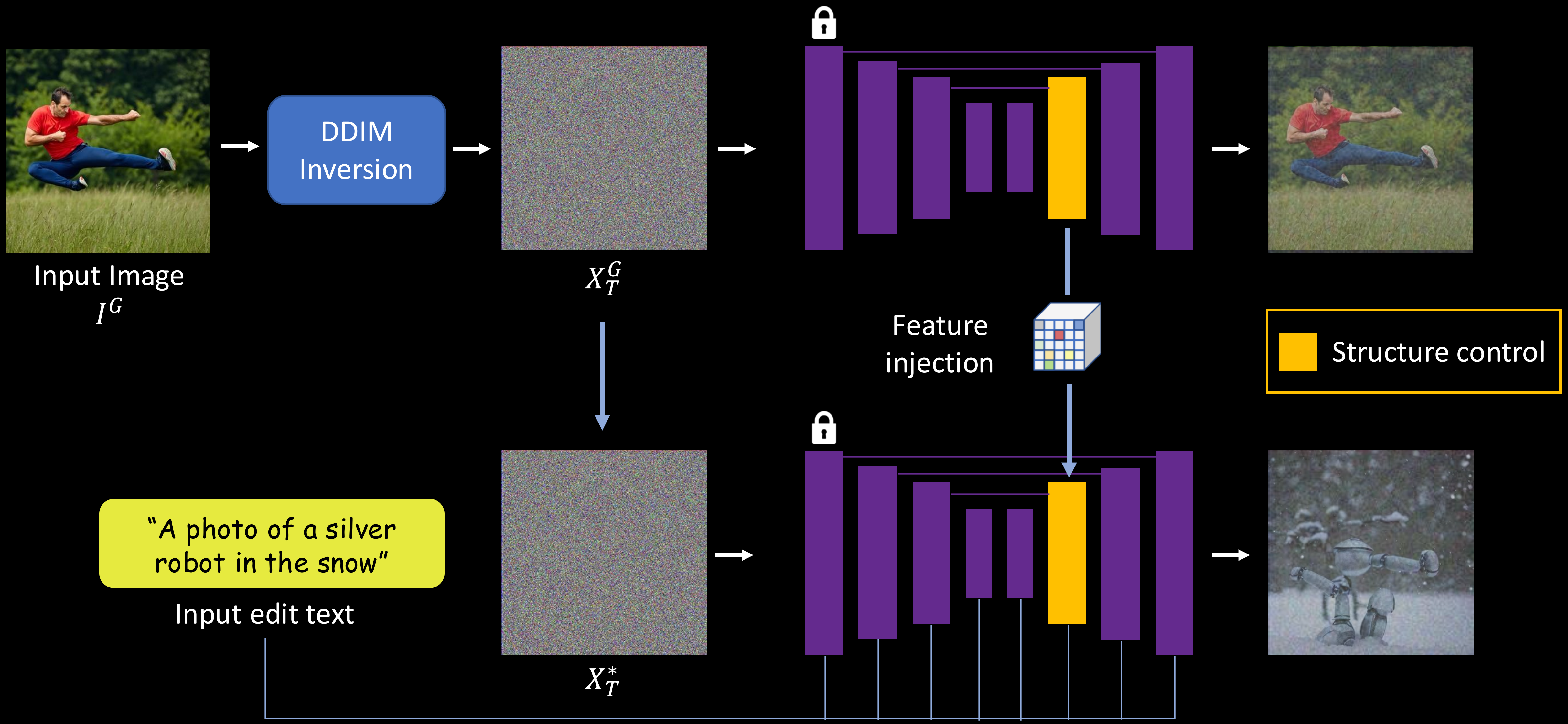
Controlling Structure in the Generation



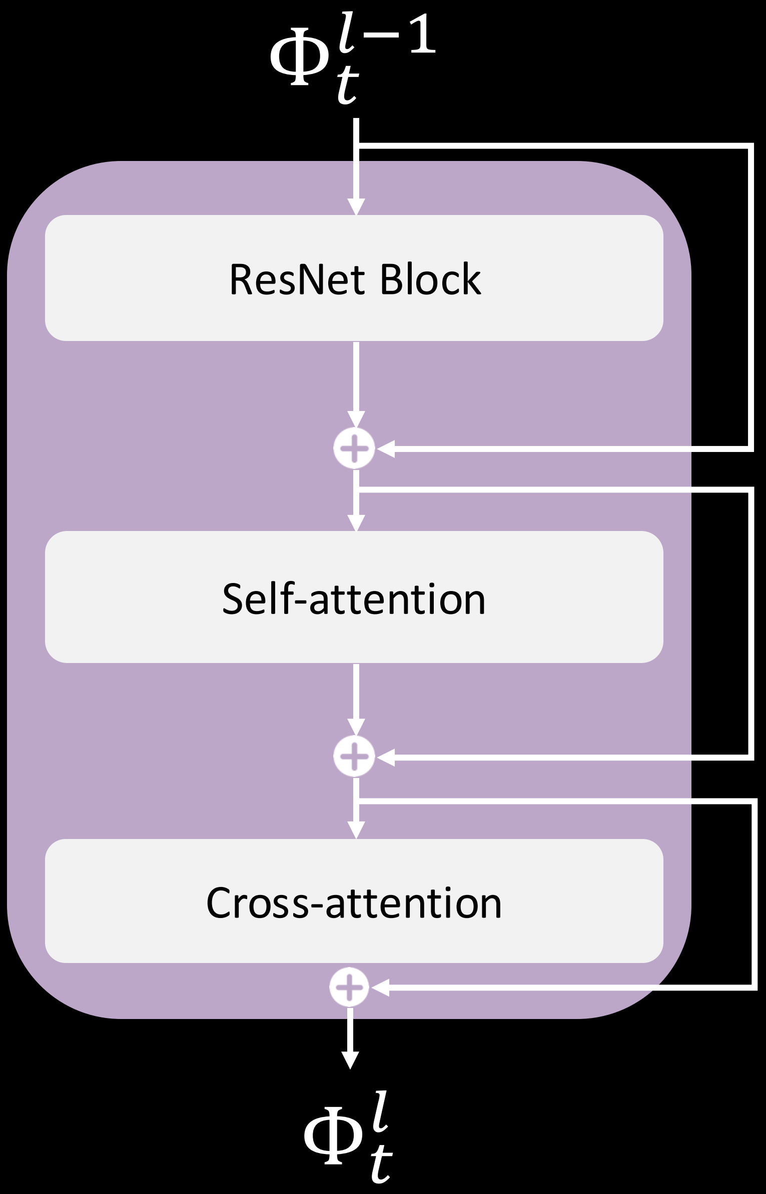
Feature Injection Result

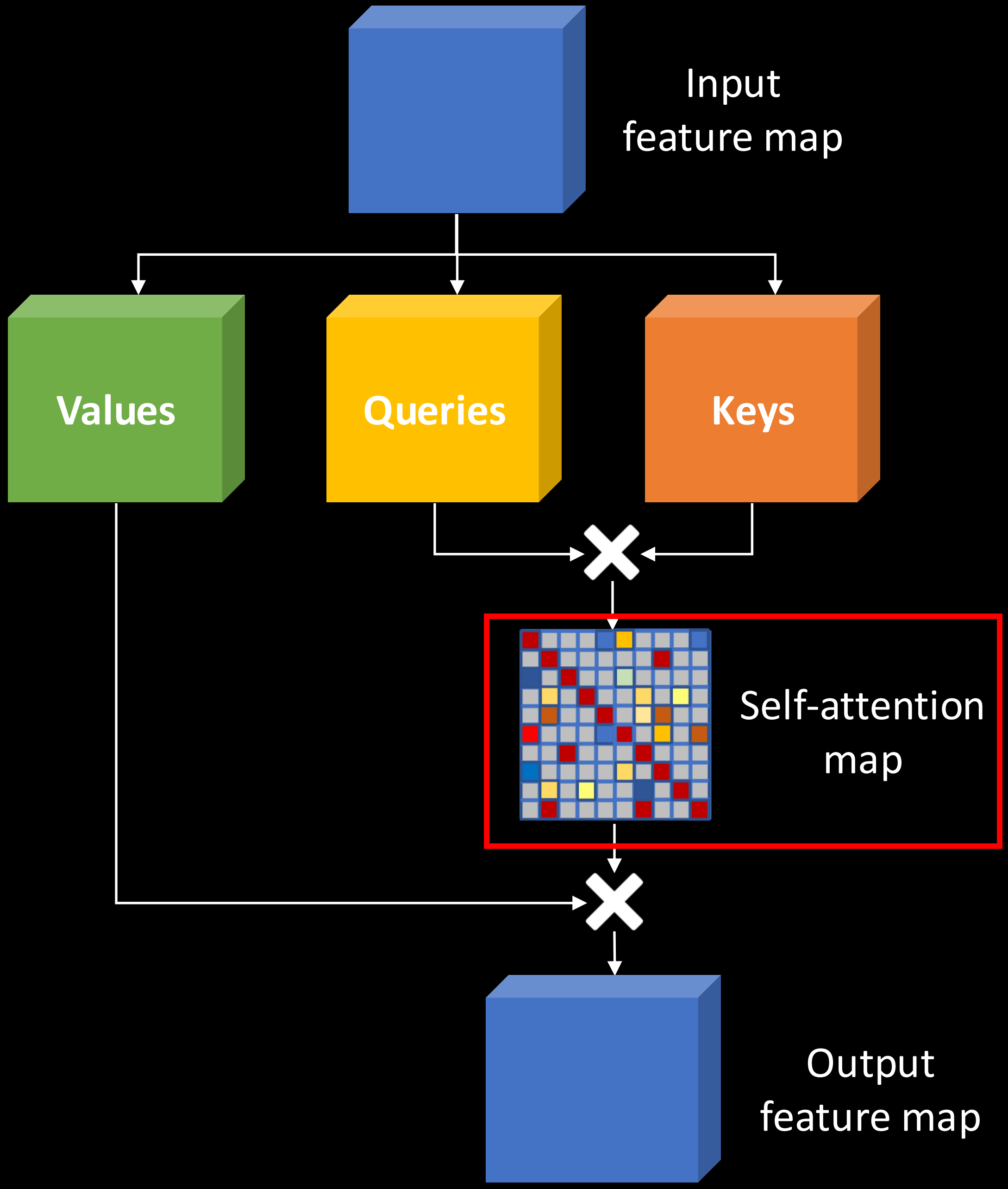
"a photo of a silver robot in the snow"

Problem with Feature Injection



Decoder Block Architecture



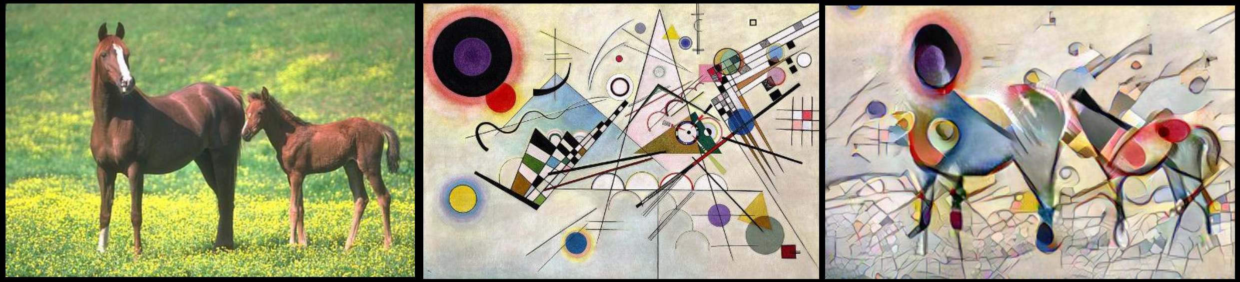


Self-Attention for Structure Control

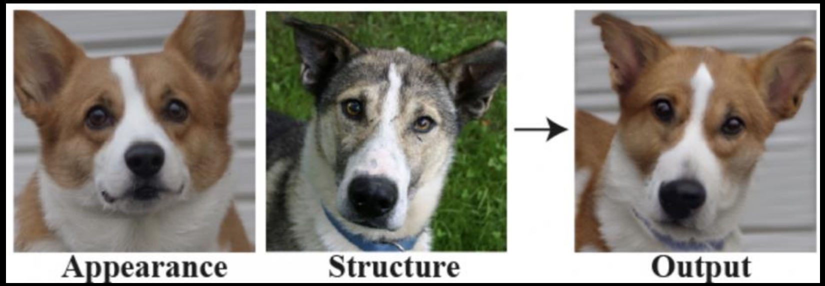
self-attention $\leftrightarrow$ self-similarity

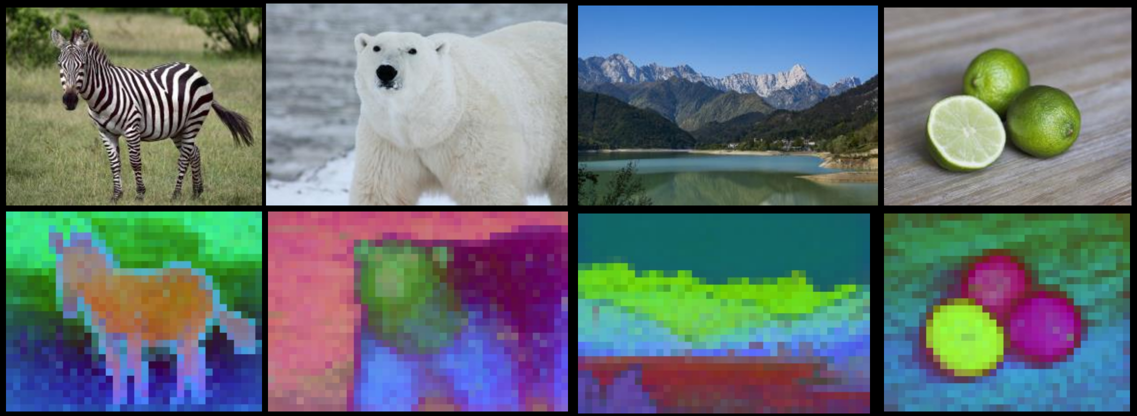
Self-similarity as a structure descriptor:

STROTSS



Splice-ViT



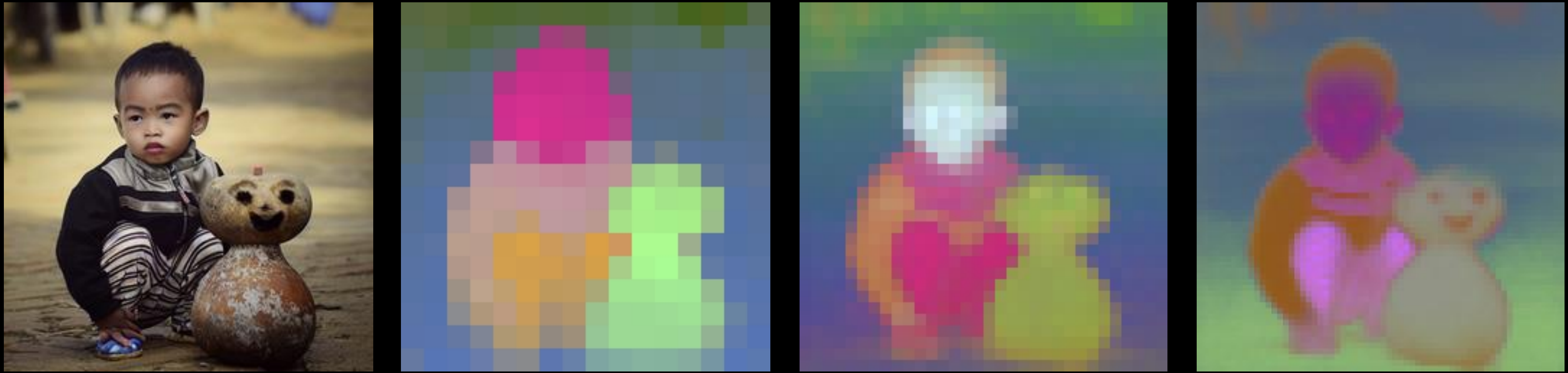


Matching Local Self-Similarities Across Images and Videos, CVPR 2007
Style Transfer by Relaxed Optimal Transport and Self-Similarity, CVPR 2019
Splicing ViT Features for Semantic Appearance Transfer, CVPR 2022

Self-Attention for Structure Control

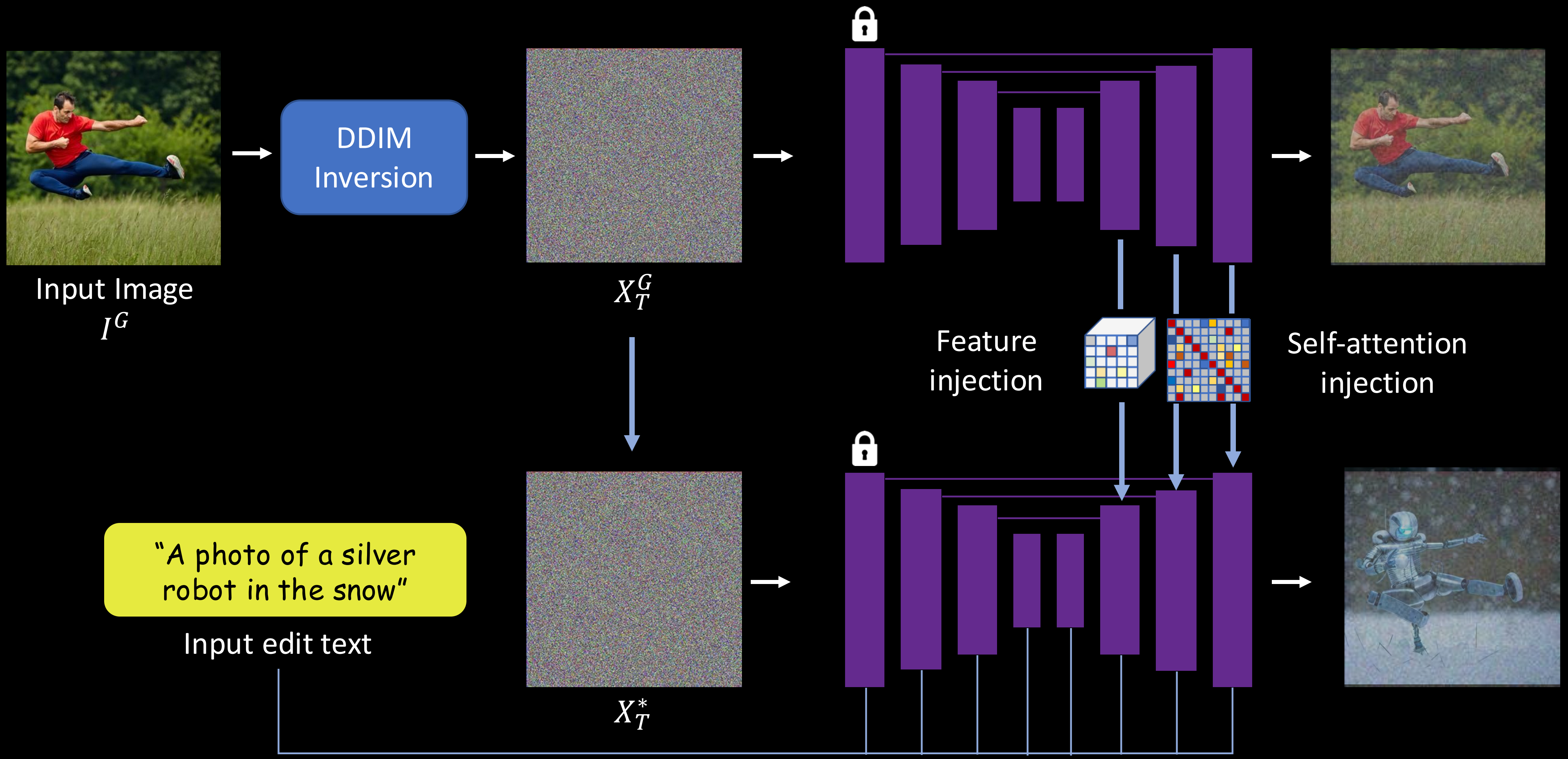
Self-attention PCA visualization:

Input image | layer = 4 | layer = 8 | layer = 11



- Self-attention aligns with the structure of the image
- Early layers align with the semantic layout
- Later layers capture higher frequencies

Plug-and-Play Diffusion Features



Feature and Self-Attention Injection Results

"a photo of a silver robot in the snow"

Feature injection
in layer 4

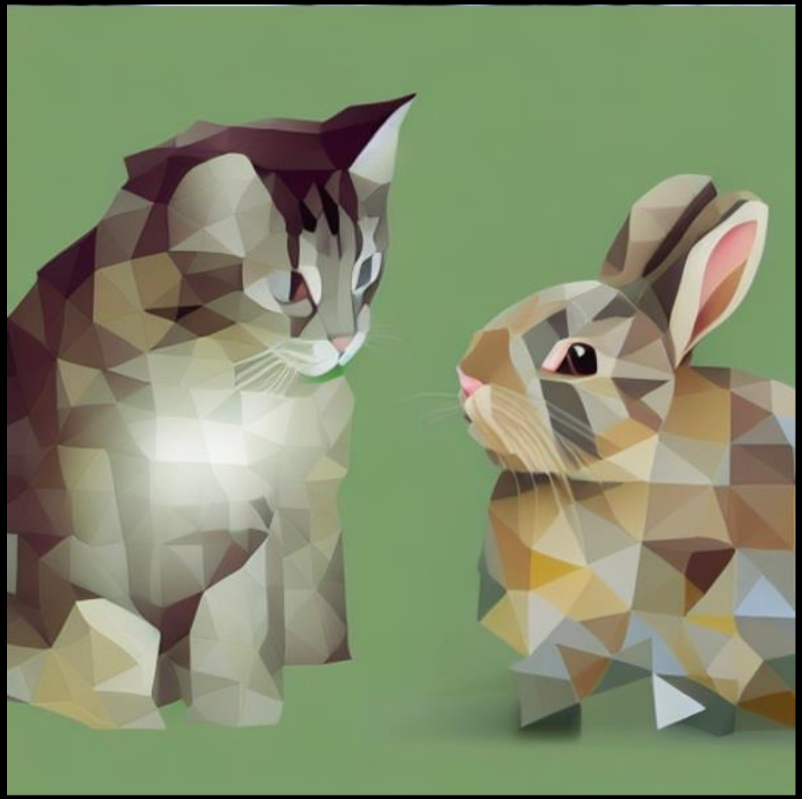
Feature injection
in layer 4
+
Self-attention injection
in layers 4-11

Results

Results

"a photorealistic image of
bear cubs in the snow"

"a polygonal illustration of
a cat and a bunny"



Results

“a wadding cake”

a skyscraper

Ablations

"A photo of a silver robot in the snow"

Layer 4 | Layers 4-8 | Layers 4-11

Feature injection

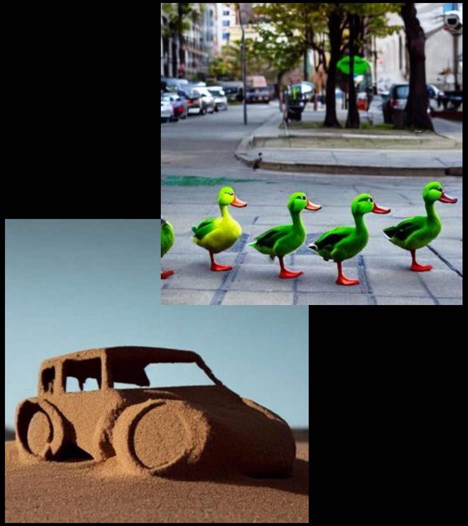
Layer 4 Features + Attention injection

Attention injection Only

Numeric evaluation

Wild TI2I: Real

Wild TI2I: Generated



- Evaluation benchmarks
  - Wild-TI2I – 148 text-image pairs

    53% real images gathered from the web.

ImageNetR

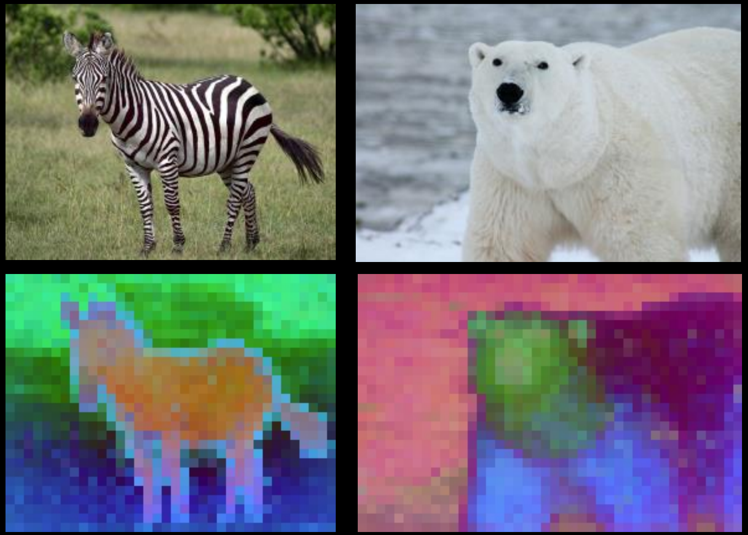
  - ImageNetR-TI2I – 150 text-image pairs.

    Various renditions of ImageNet object classes.

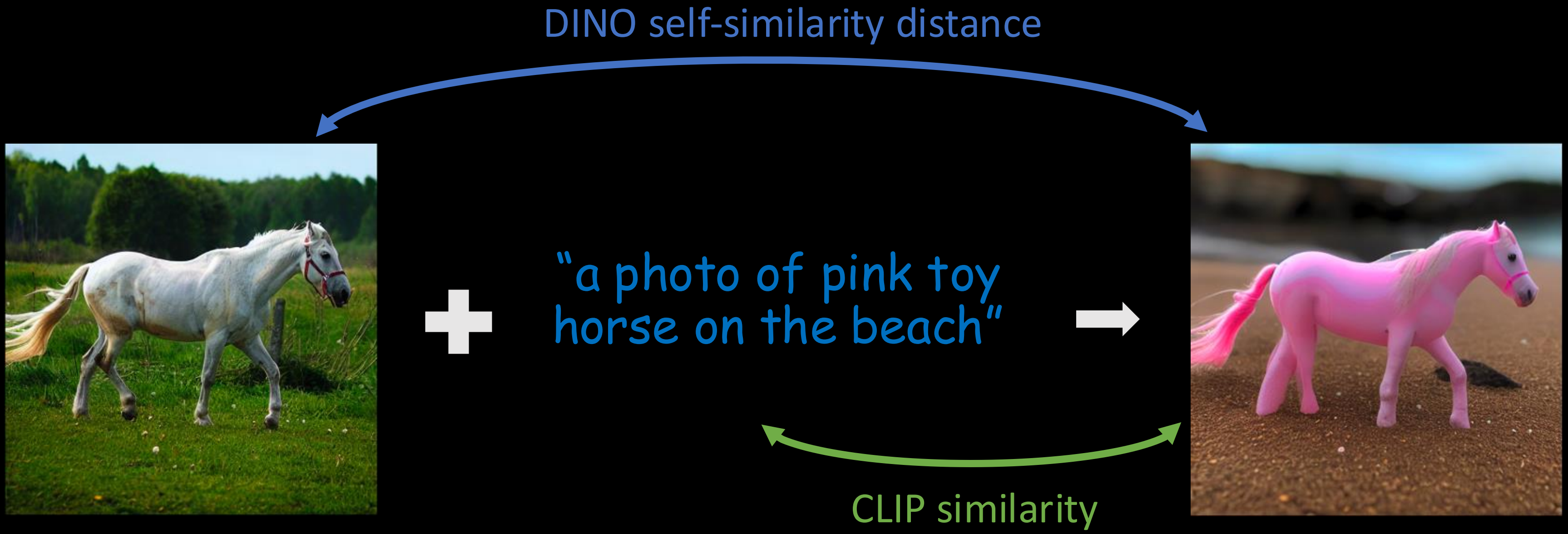
The many faces of robustness: A critical analysis of out-of-distribution generalization, ICCV 2021

Numeric evaluation

- Evaluation metrics
  - Structure preservation
    - DINO keys SSIM distance (lower is better)
  - Translation prompt fidelity
    - CLIP score (higher is better)

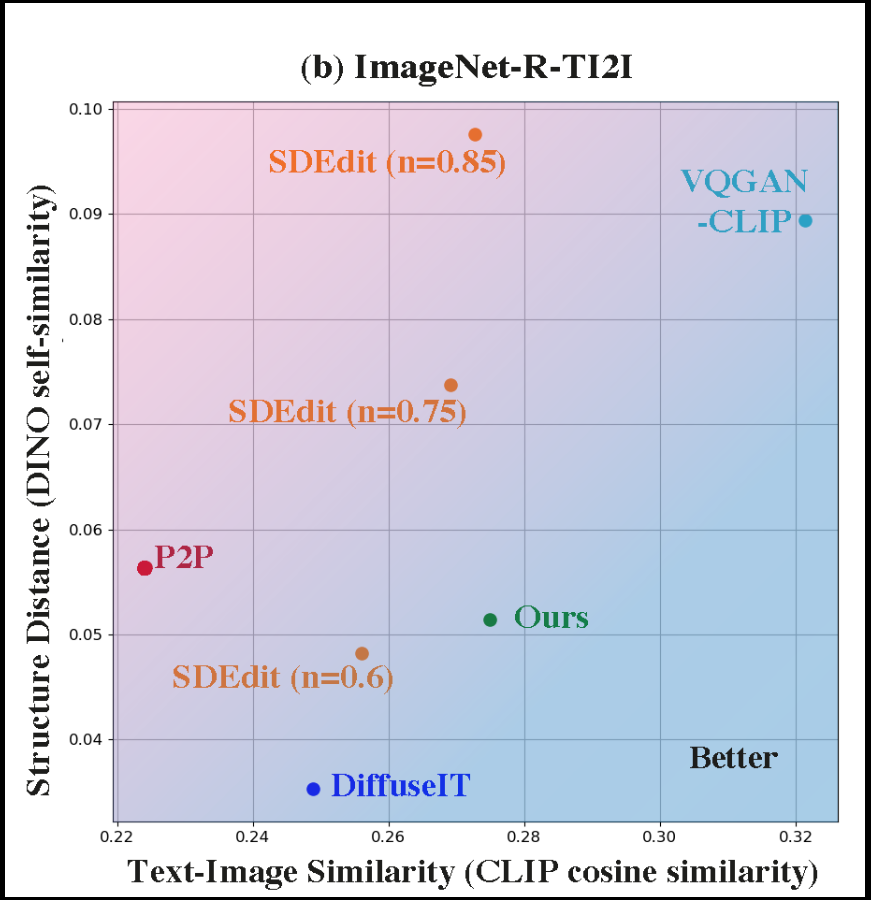


PCA visualization of DINO keys self-similarity



Splicing ViT Features for Semantic Appearance Transfer, CVPR 2022

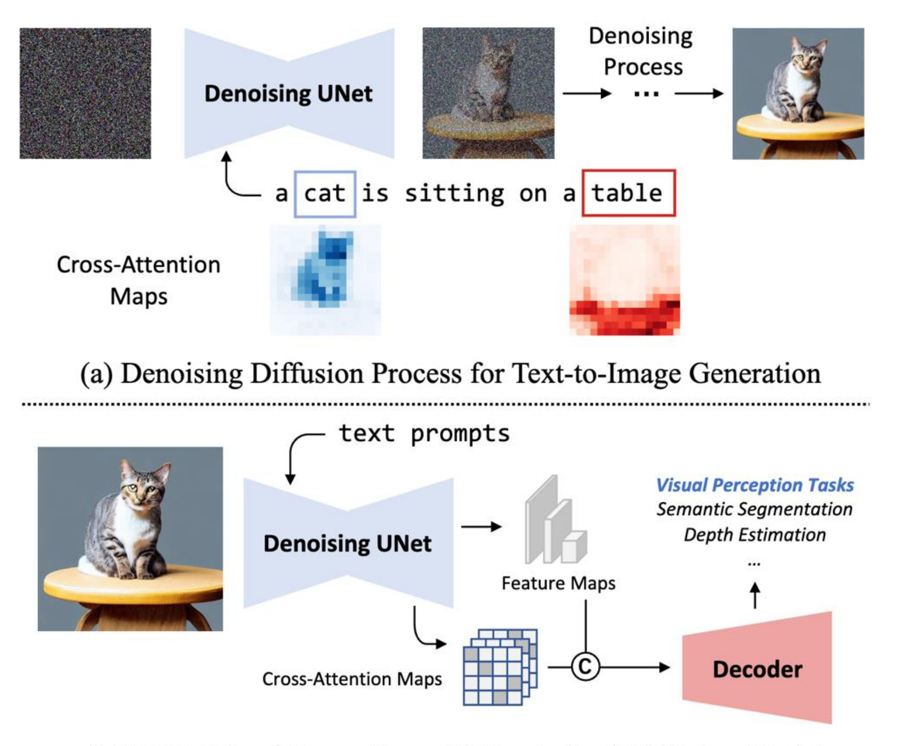
Numeric evaluation



What else can we do based on the diffusion features?

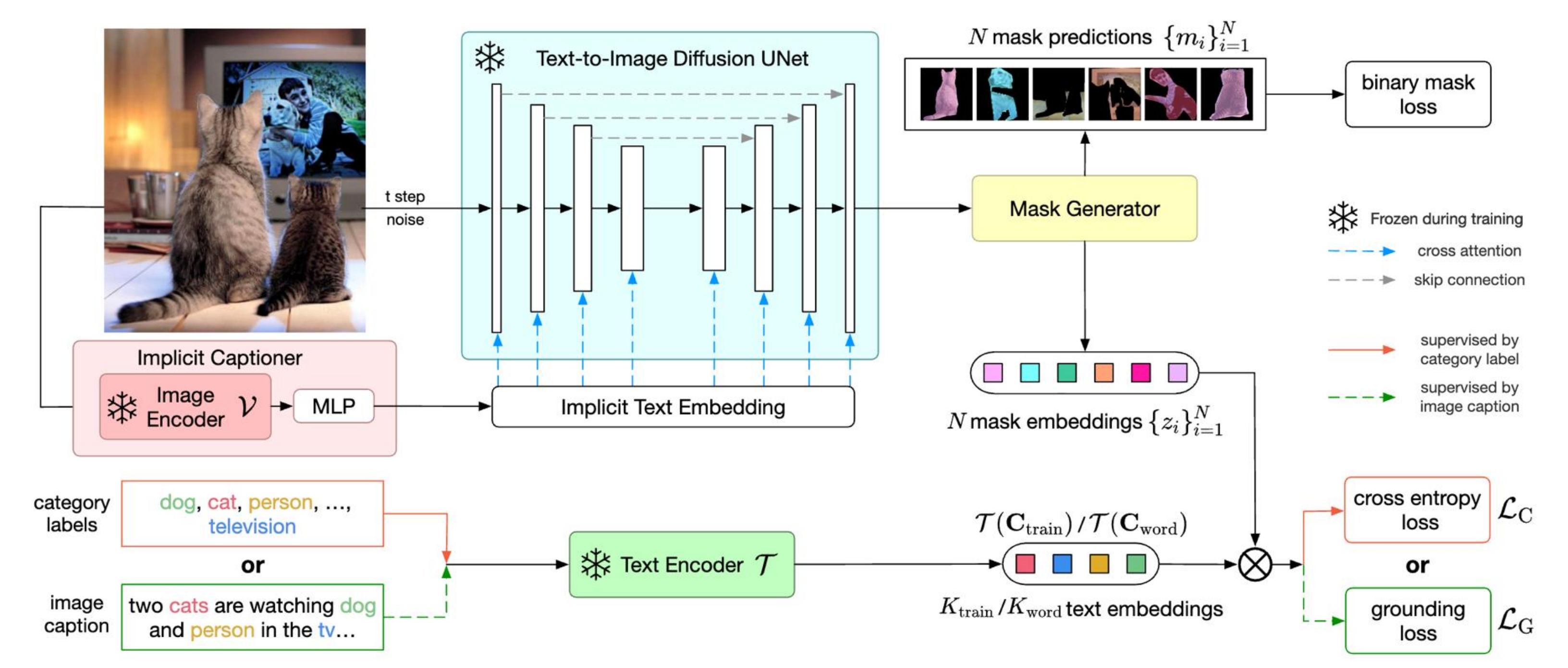
Diffusion features for other perception tasks

- Depth estimation, semantic segmentation, panoptic segmentation



(a) Denoising Diffusion Process for Text-to-Image Generation

(b) **VPD**: Visual Perception with Pre-trained Diffusion Models

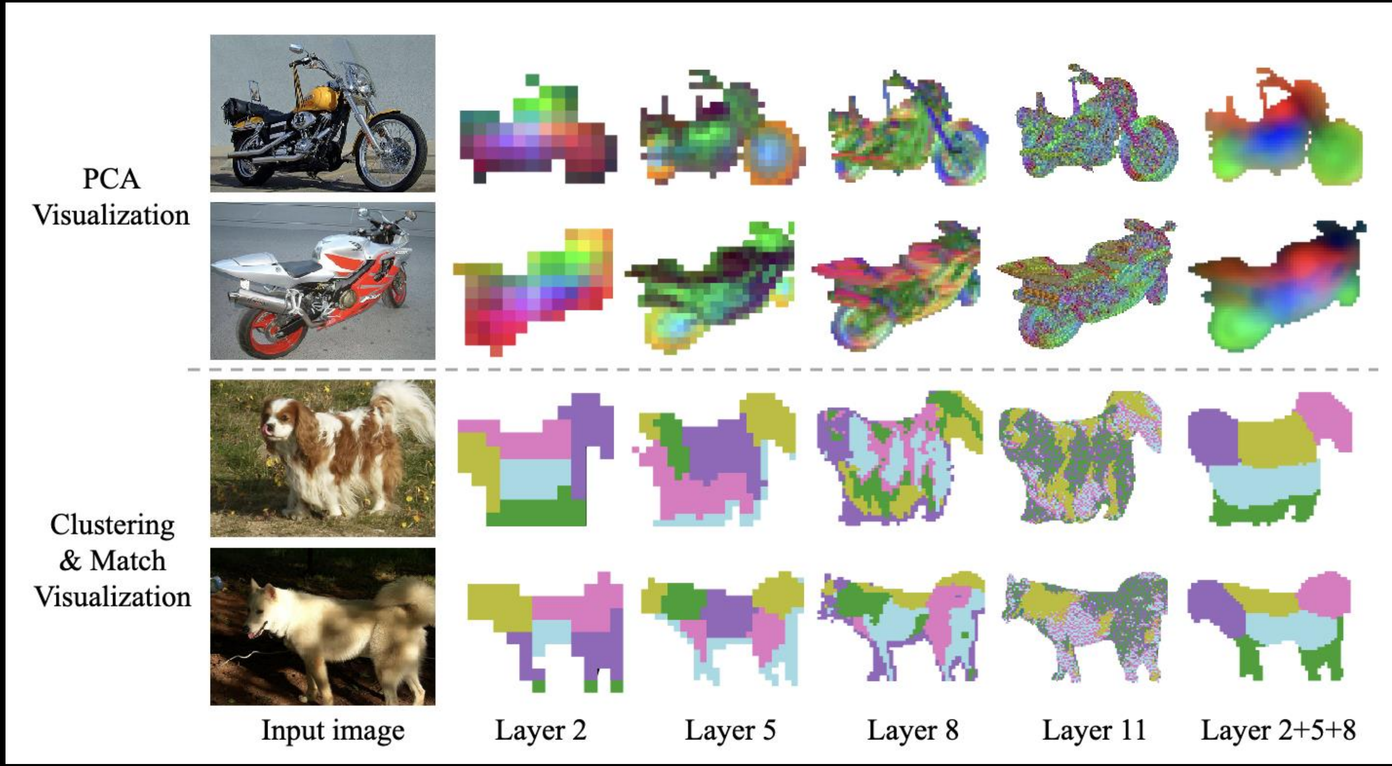


Unleashing Text-to-Image Diffusion Models for Visual Perception, ICCV23

Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR23

Diffusion features for other perception tasks

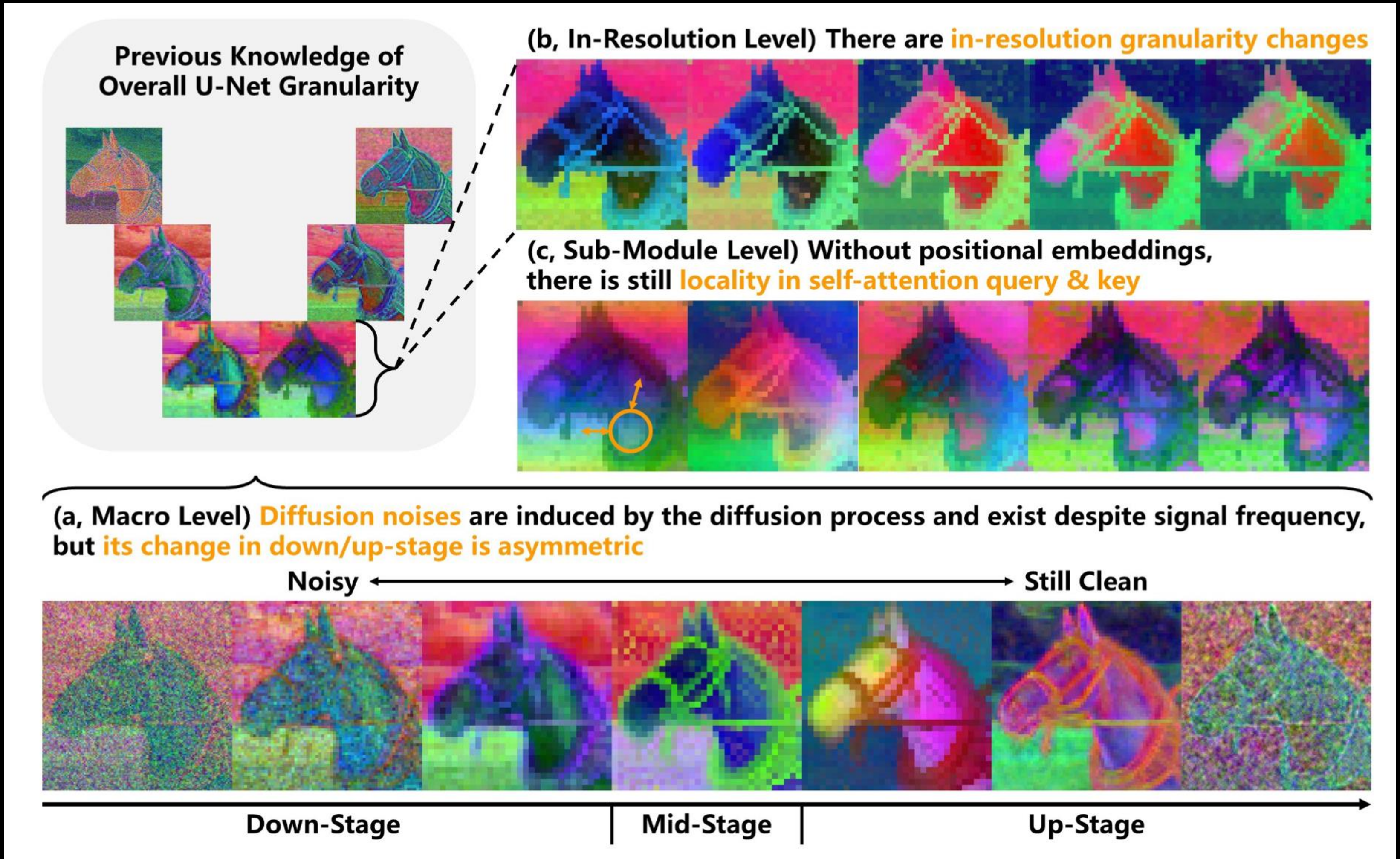
- Feature correspondence



	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
S	SCOT [34]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20.0	42.9	42.5	31.1	29.8	35.0	27.7	24.4	48.4	40.8	35.6
	CATs* [9]	52.0	34.7	72.2	34.3	49.9	57.5	43.6	66.5	24.4	63.2	56.5	52.0	42.6	41.7	43.0	33.6	72.6	58.0	49.9
	PMNC* [30]	54.1	35.9	74.9	36.5	42.1	48.8	40.0	72.6	21.1	67.6	58.1	50.5	40.1	54.1	43.3	35.7	74.5	59.9	50.4
	SCorrSAN* [24]	57.1	40.3	78.3	38.1	51.8	57.8	47.1	67.9	25.2	71.3	63.9	49.3	45.3	49.8	48.8	40.3	77.7	69.7	55.3
	CATs++* [10]	60.6	46.9	82.5	41.6	56.8	64.9	50.4	72.8	29.2	75.8	65.4	62.5	50.9	56.1	54.8	48.2	80.9	**74.9**	59.9
	DINOv2-ViT-B/14†	80.4	60.2	88.1	59.5	54.9	82.0	73.5	89.1	53.3	85.5	73.6	73.8	65.2	72.3	43.6	65.6	91.4	60.3	69.9
	Stable Diffusion† (**Ours**)	75.6	60.3	87.3	41.5	50.8	68.4	77.2	81.4	44.3	79.4	62.8	67.7	64.9	71.6	57.8	53.3	89.2	65.1	66.3
	Fuse-ViT-B/14† (**Ours**)	**81.2**	**66.9**	**91.6**	**61.4**	**57.4**	**85.3**	**83.1**	**90.8**	**54.5**	**88.5**	**75.1**	**80.2**	**71.9**	**77.9**	**60.7**	**68.9**	**92.4**	65.8	**74.6**
G	GANgealing [42]	-	37.5	-	-	-	-	-	67.0	-	-	23.1	-	-	-	-	-	-	57.9	-
U^T	VGG+MLS [1]	29.5	22.7	61.9	26.5	20.6	25.4	14.1	23.7	14.2	27.6	30.0	29.1	24.7	27.4	19.1	19.3	24.4	22.6	27.4
	DINO+MLS [1, 5]	49.7	20.9	63.9	19.1	32.5	27.6	22.4	48.9	14.0	36.9	39.0	30.1	21.7	41.1	17.1	18.1	35.9	21.4	31.1
	NeuCongeal [39]	-	29.1	-	-	-	-	-	53.3	-	-	35.2	-	-	-	-	-	-	-	-
	ASIC [18]	57.9	25.2	68.1	24.7	35.4	28.4	30.9	54.8	21.6	45.0	47.2	39.9	26.2	48.8	14.5	24.5	49.0	24.6	36.9
U^N	DINOv1-ViT-S/8 [2]	57.2	24.1	67.4	24.5	26.8	29.0	27.1	52.1	15.7	42.4	43.3	30.1	23.2	40.7	16.6	24.1	31.0	24.9	33.3
	DINOv2-ViT-B/14	72.7	62.0	85.2	**41.3**	40.4	52.3	51.5	71.1	36.2	67.1	64.6	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.6
	Stable Diffusion (**Ours**)	63.1	55.6	80.2	33.8	44.9	49.3	47.8	74.4	38.4	70.8	53.7	61.1	54.4	55.0	54.8	53.5	65.0	53.3	57.2
	Fuse-ViT-B/14 (**Ours**)	**73.0**	**64.1**	**86.4**	40.7	**52.9**	**55.0**	**53.8**	**78.6**	**45.5**	**77.3**	**64.7**	**69.7**	**63.3**	**69.2**	**58.4**	**67.6**	**66.2**	**53.5**	**64.0**

A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence, NeurIPS23

More recent analysis on diffusion features



Not All Diffusion Model Activations Have Been Evaluated as Discriminative Features, NeurIPS24

Questions

- While the paper demonstrates the effectiveness of manipulating cross-attention maps to control image generation. How might direct manipulation of other components enable different types of semantic control? Are there other places to edit apart from the low resolution cross attention maps?
- While prompt-2-prompt is a good paper, I don't believe modifying X-attention layers can solve any editing problems, such as removing a specific object rather than having global editing. I was wondering if this believe is still valid in late 2024 or something changed in the past 6 months?