Diffusion Models for Visual Content Creation

Niloy Mitra, Duygu Ceylan, Paul Guerrero,

Daniel Cohen-Or, Or Patashnik, Chun-Hao Huang, Minhyuk Sung

Part 3: The Power of Attention Layers

UCL Adobe TEL AVIV UNIVERSITY KAIST

https://geometry.cs.ucl.ac.uk/courses/diffusion4ContentCreation_sigg24/

Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

Break

Personalization and Editing

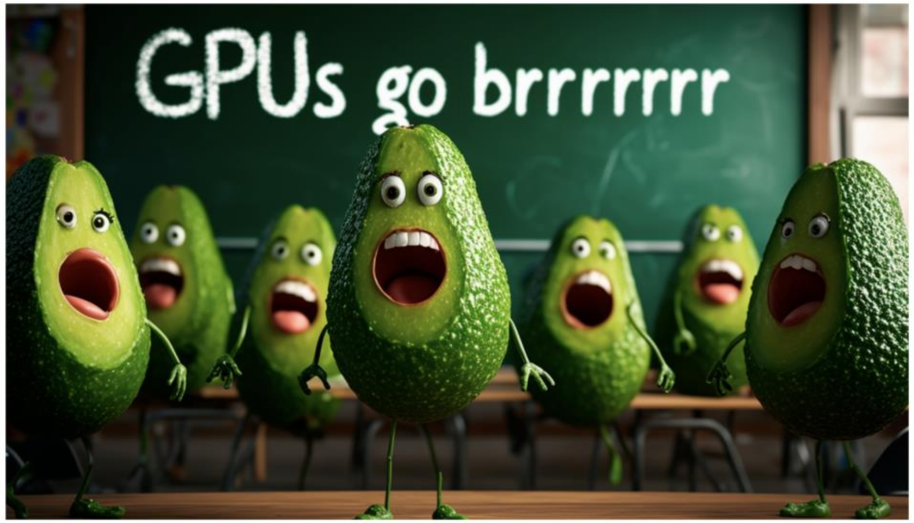
Beyond Single (RGB) Image Generation

Diffusion Models for 3D Generation

How Did We Get From

Here

To Here



Prompt: A surreal and humorous scene in a classroom with the words 'GPUs go brrrrrr' written in white chalk on a blackboard. In front of the blackboard, a group of students are celebrating. These students are uniquely depicted as avocados, complete with little arms and legs, and faces showing expressions of joy and excitement. The scene captures a playful and imaginative atmosphere, blending the concept of a traditional classroom with the whimsical portrayal of avocado students.

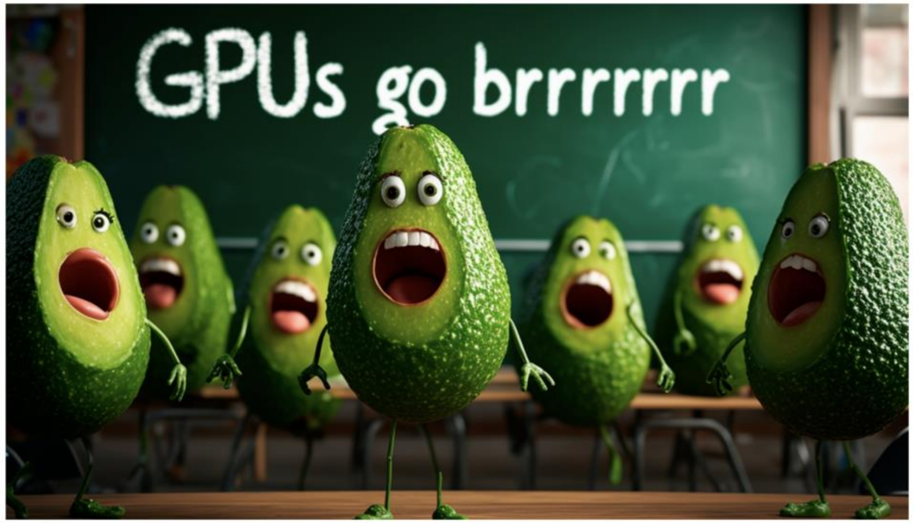
Two cats doing research.

Isometric underwater Atlantis city with a Greek temple in a bubble.

DALL·E 3

A 2D animation of a folk music band composed of anthropomorphic autumn leaves, each playing traditional bluegrass instruments, amidst a rustic forest setting dappled with the soft light of a harvest moon.

To Here



Prompt: A surreal and humorous scene in a classroom with the words 'GPUs go brrrrrr' written in white chalk on a blackboard. In front of the blackboard, a group of students are celebrating. These students are uniquely depicted as avocados, complete with little arms and legs, and faces showing expressions of joy and excitement. The scene captures a playful and imaginative atmosphere, blending the concept of a traditional classroom with the whimsical portrayal of avocado students.

Two cats doing research.

Isometric underwater Atlantis city with a Greek temple in a bubble.

DALL·E 3

A 2D animation of a folk music band composed of anthropomorphic autumn leaves, each playing traditional bluegrass instruments, amidst a rustic forest setting dappled with the soft light of a harvest moon.

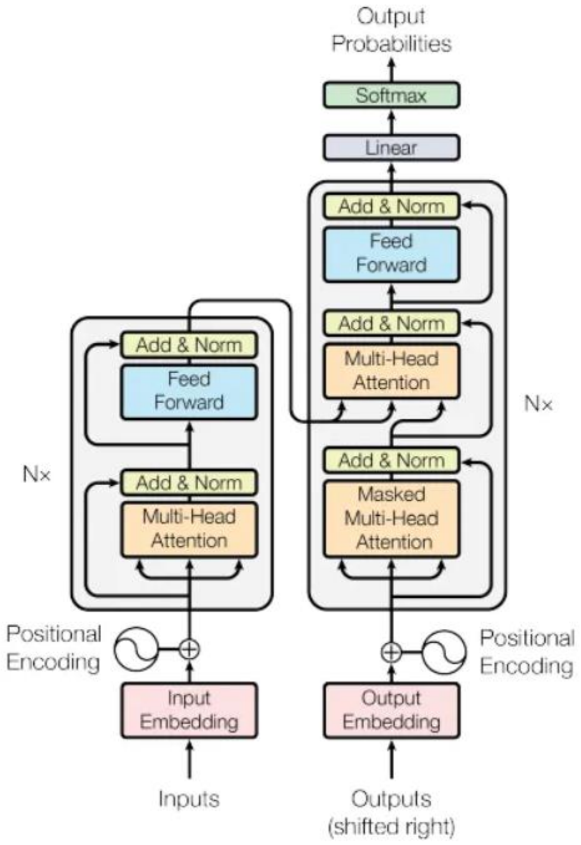
Scale, data, ...

Common to all these models is the use of attention layers

In other words:

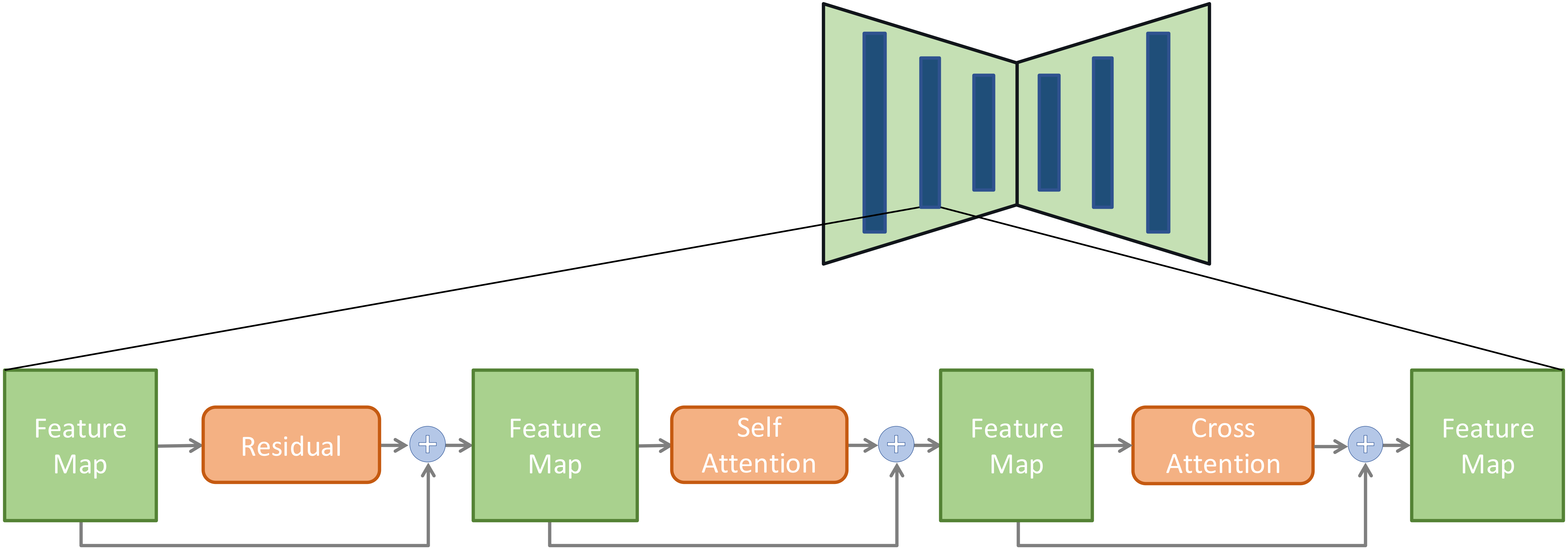
“Attention is all you need” [Vaswani et al. 2017]

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

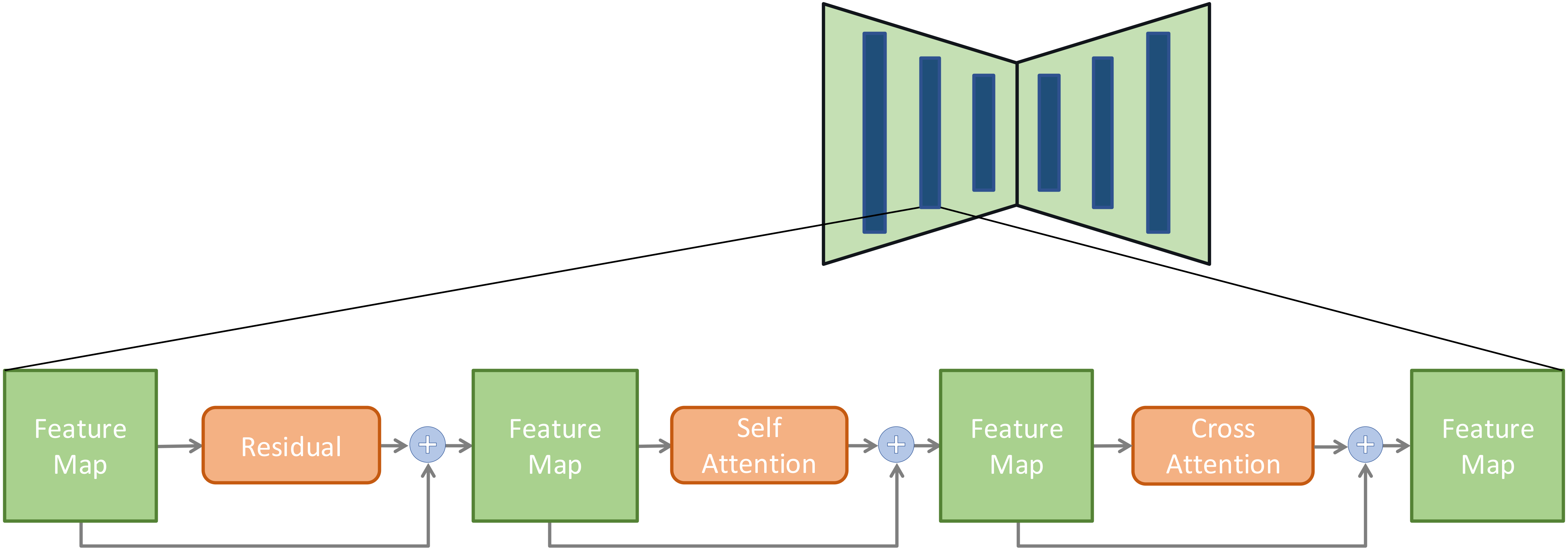


[Attention Is All You Need, Vaswani et al., NeurIPS 2017]

Stable Diffusion’s Model (UNet)

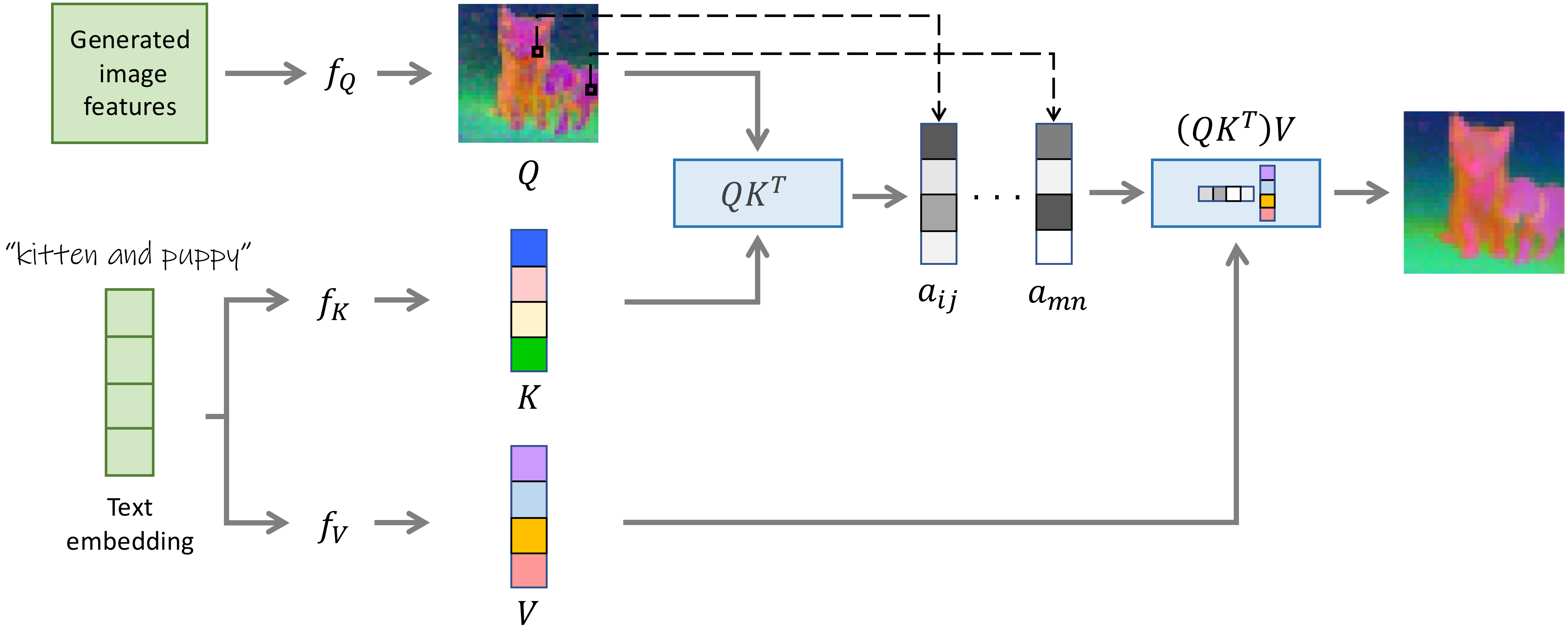


Stable Diffusion's Model (UNet)



Key for text conditioning

Cross-Attention Layers



Generated image features

f_Q

Q

QK^T

a_{ij}

a_{mn}

$(QK^T)V$

“kitten and puppy”

f_K

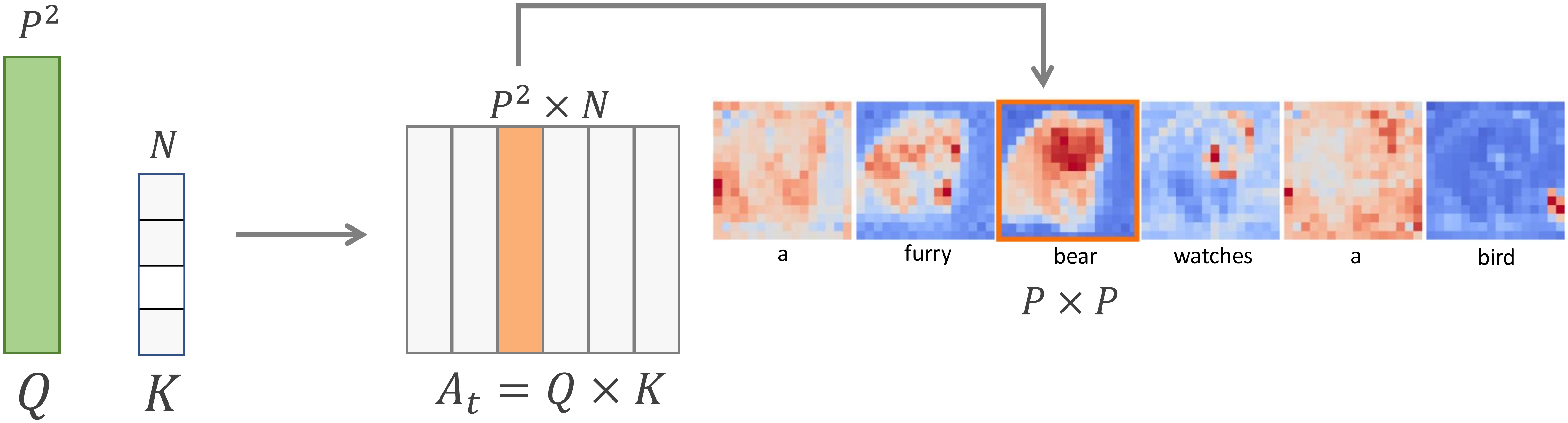
K

Text embedding

f_V

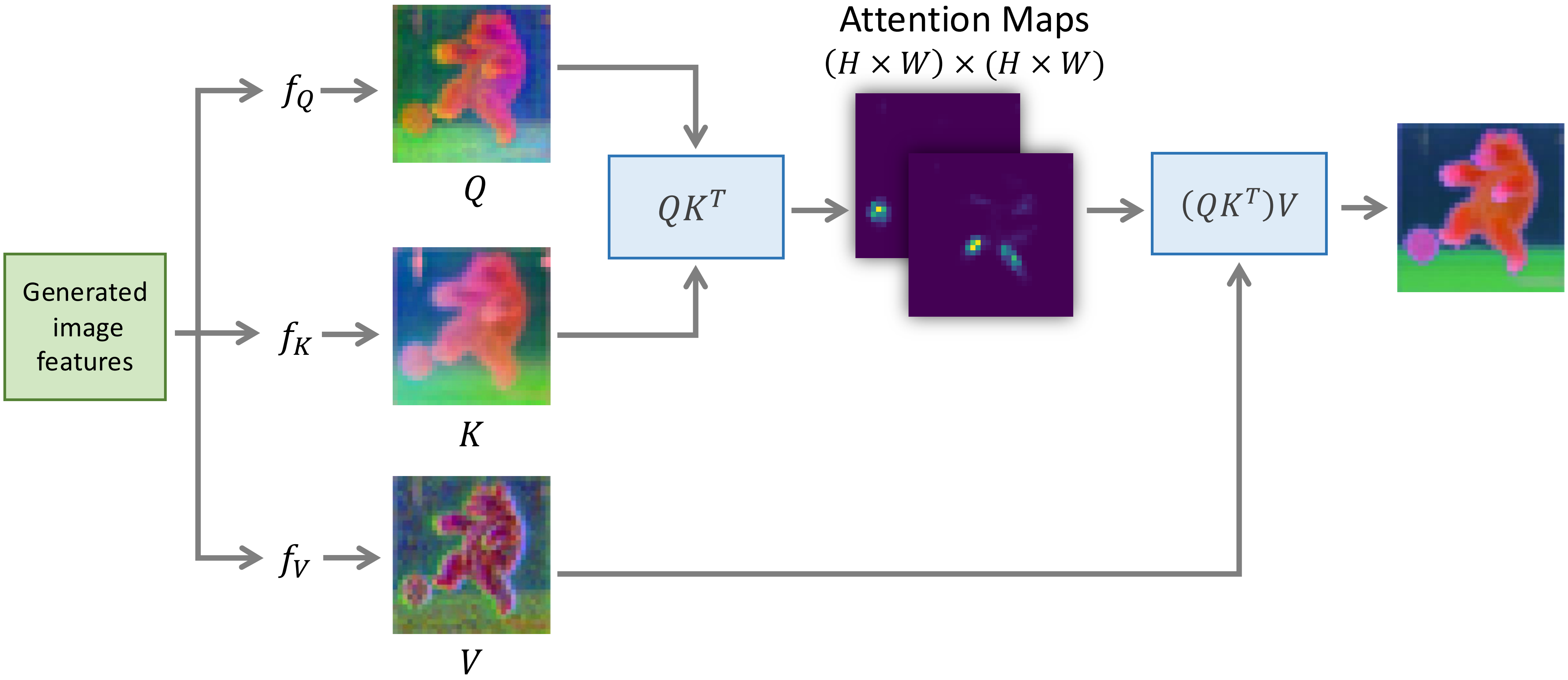
V

Cross-Attention Layers Another Point of View

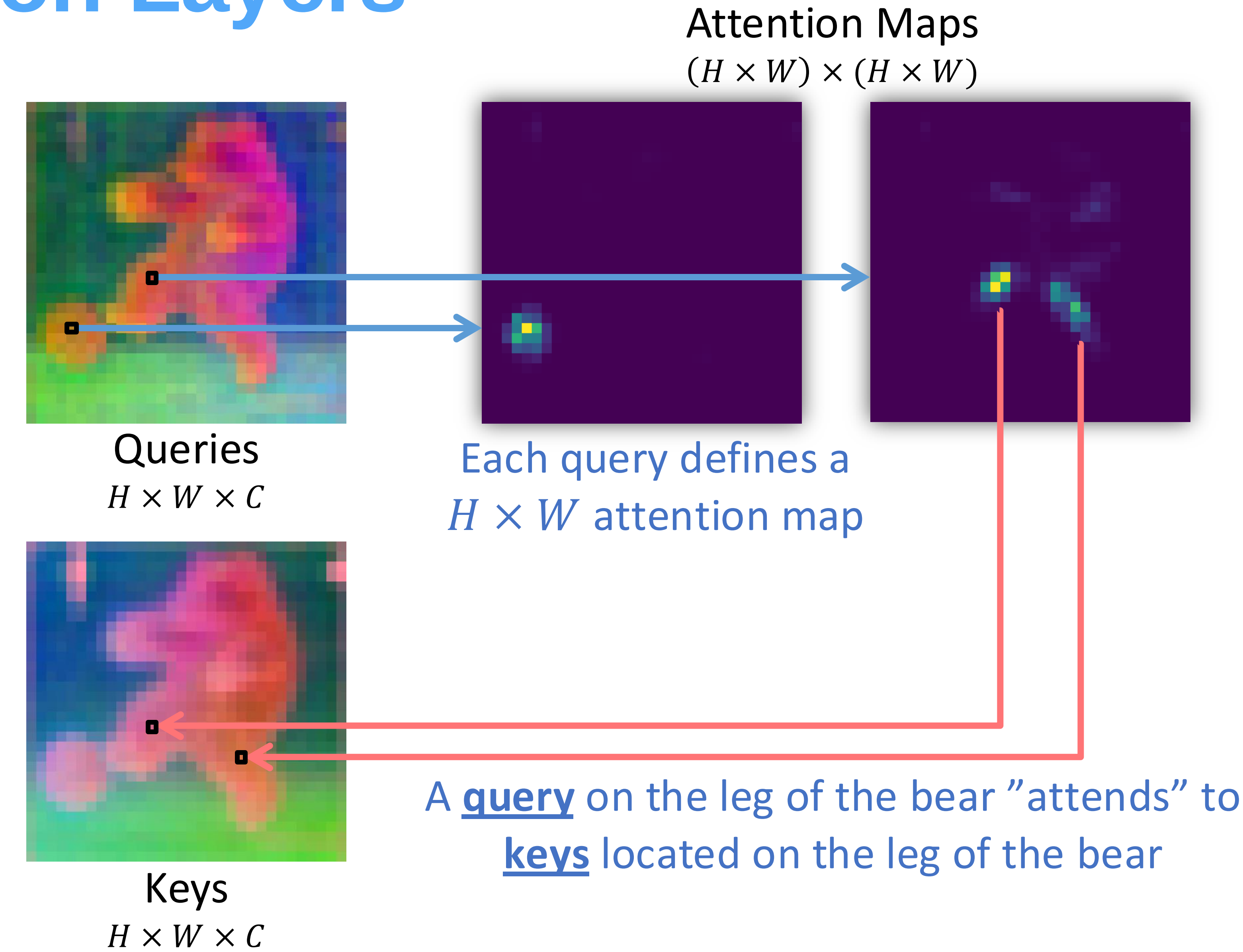


Text Embedding

Self-Attention Layers



Self-Attention Layers



Attention Maps

$(H \times W) \times (H \times W)$

Queries

$H \times W \times C$

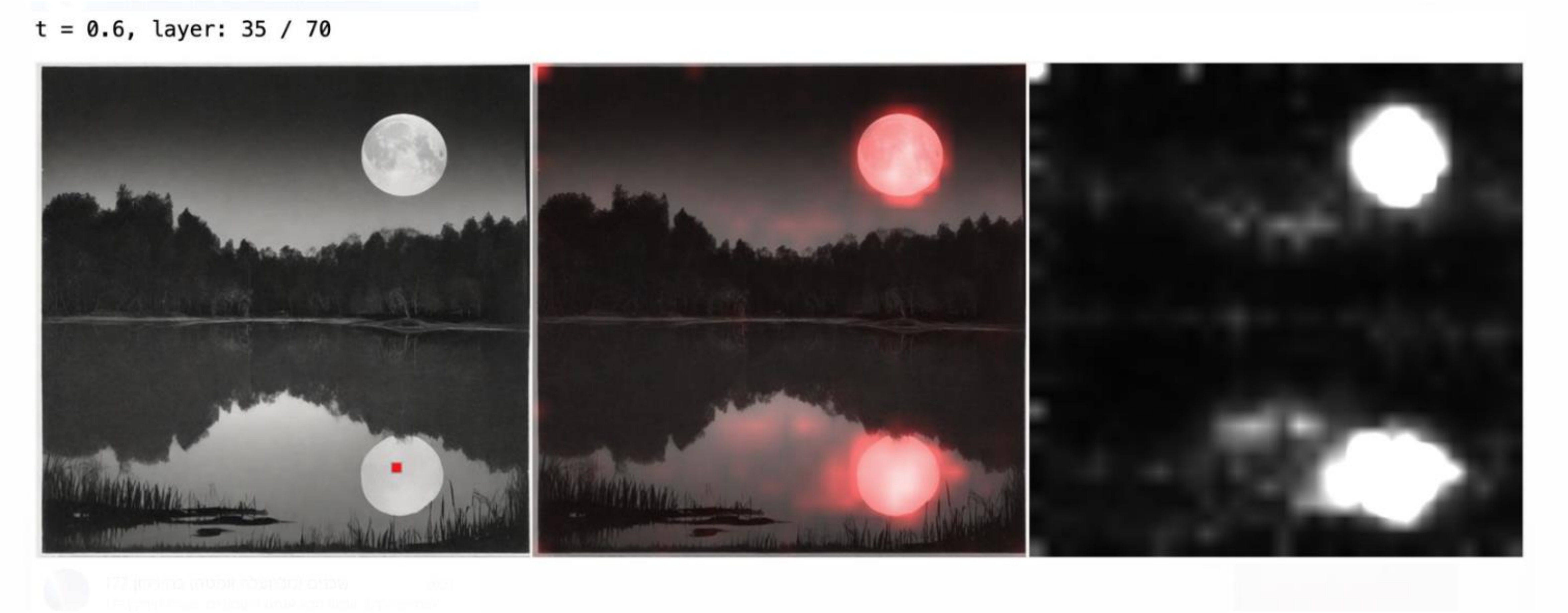
Each query defines a $H \times W$ attention map

Keys

$H \times W \times C$

A **query** on the leg of the bear ”attends” to **keys** located on the leg of the bear

Self-Attention Layers



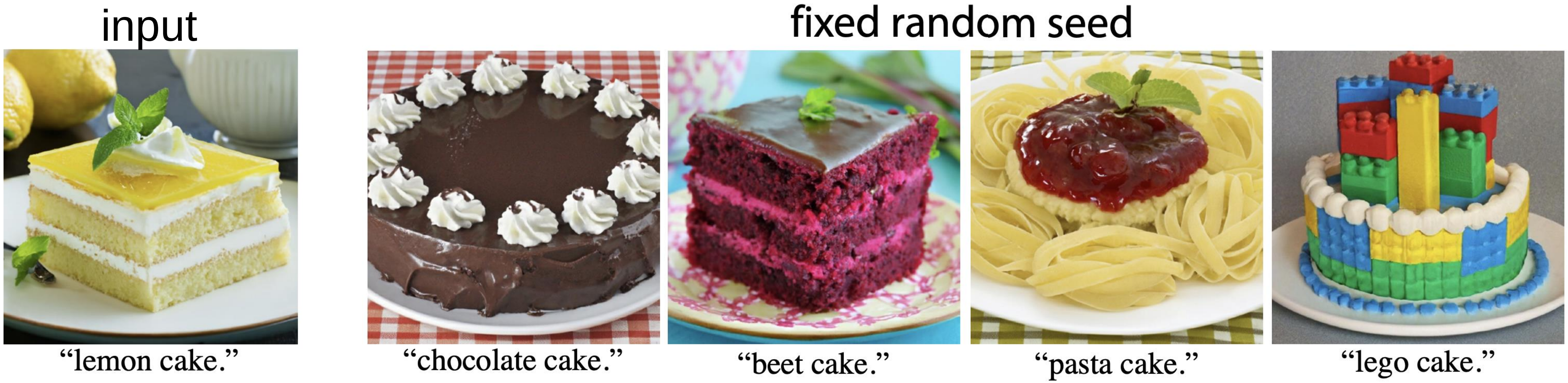
t = 0.6, layer: 35 / 70

Semantics in Attention Layers

Attention-Based Text Guided Image Editing in Diffusion Models

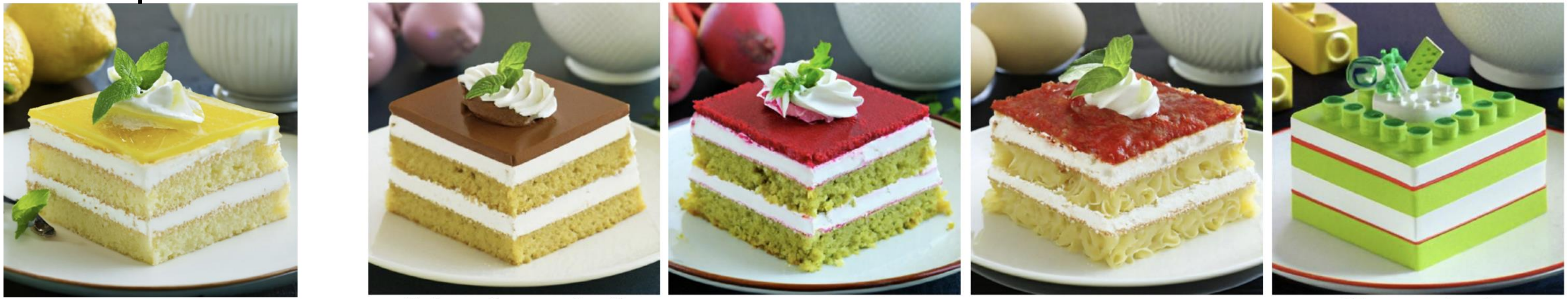
Prompt-to-Prompt [Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D., ICLR 2023]
Plug-and-Play features [Tumanyan et al., CVPR 2023]
Null-text Inversion [Mokady et al., CVPR 2023]
pix2pix-zero [Parmar et al., SIGGRAPH 2023]
MasaCtrl [Cao et al., ICCV 2023]
Rich-text Editing [Ge et al., ICCV 2023]
Self-Guidance [Epstein et al., NeurIPS 2023]
Directed Diffusion [Ma et al., 2023]

Editing an Image with Text Prompt



Editing an Image with Text Prompt

input



"lemon cake."

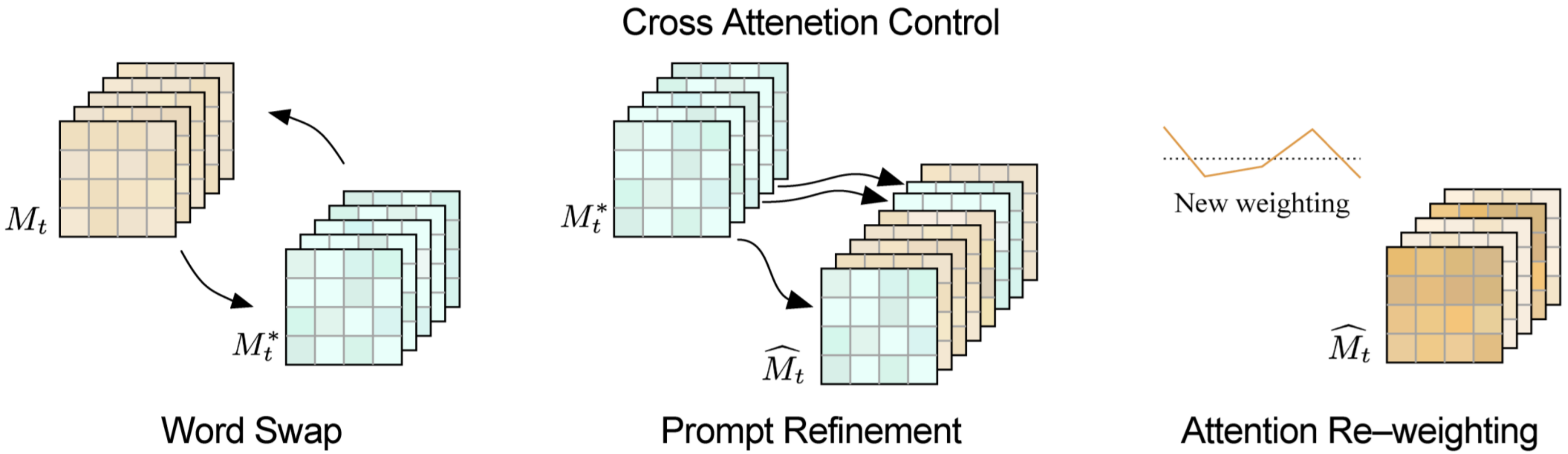
"chocolate cake."

"beet cake."

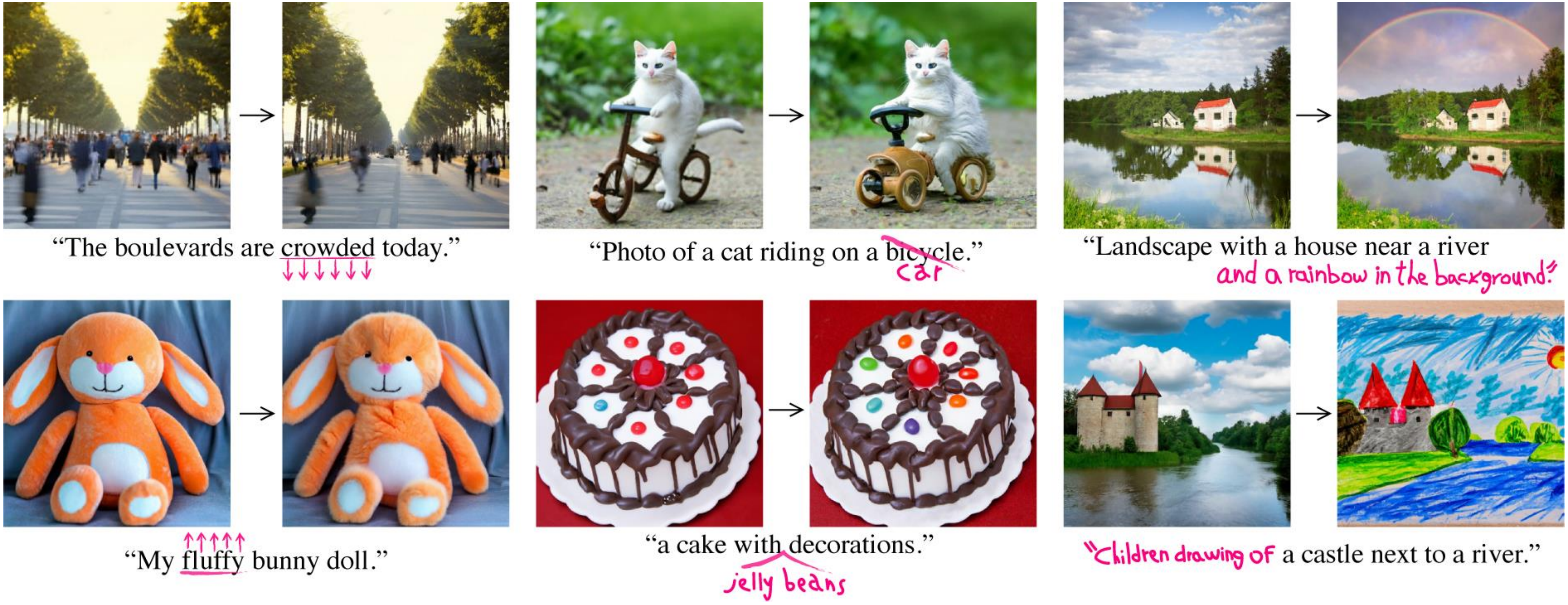
"pasta cake."

"lego cake."

Injecting Attention Maps



Prompt-to-Prompt Results



"The boulevards are crowded today."
↓↓↓↓↓↓

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

"Landscape with a house near a river
and a rainbow in the background!"

↑↑↑↑↑
"My fluffy bunny doll."

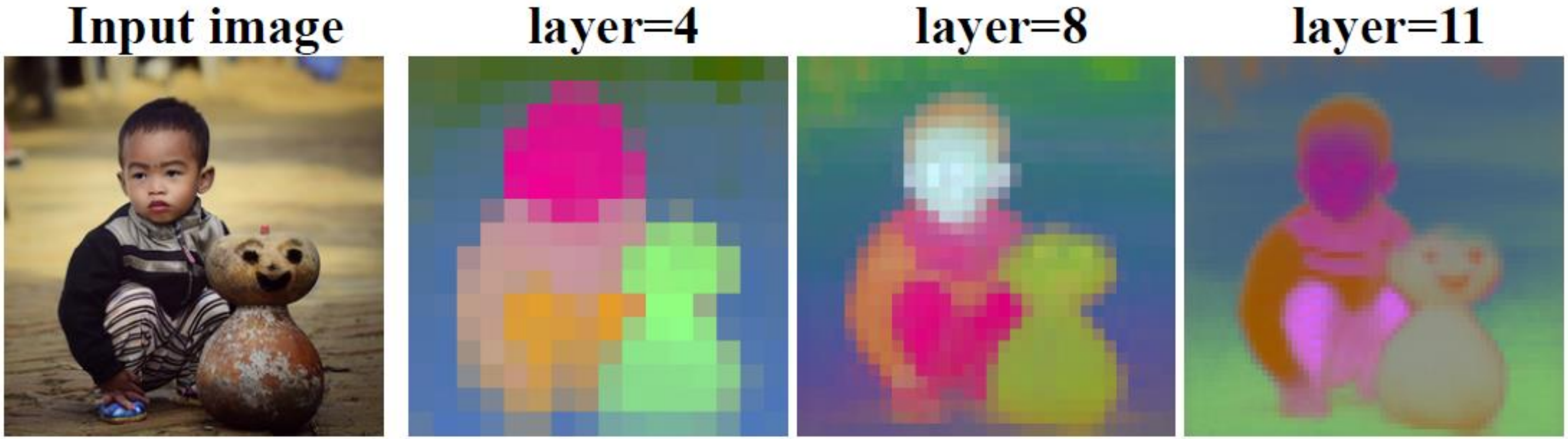
"a cake with jelly beans decorations."

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

Segmentation

Localizing Object-level Shape Variations [Patashnik, O., Garibi, D. Azuri, I., Elor, H., Cohen-Or, D. ICCV 2023]
Label-efficient semantic segmentation with diffusion models [Baranchuk et al., ICLR 2022]
Text-Guided Synthesis of Eulerian Cinemagraphs [Mahapatra et al., SIGGRAPH Asia 2023]
SLiMe [Khani et al., ICLR 2024]
EmerDiff [Namekata et al., ICLR 2024]
LIME [Simsar et al., 2023]
From Text to Mask [Xiao et al., 2023]

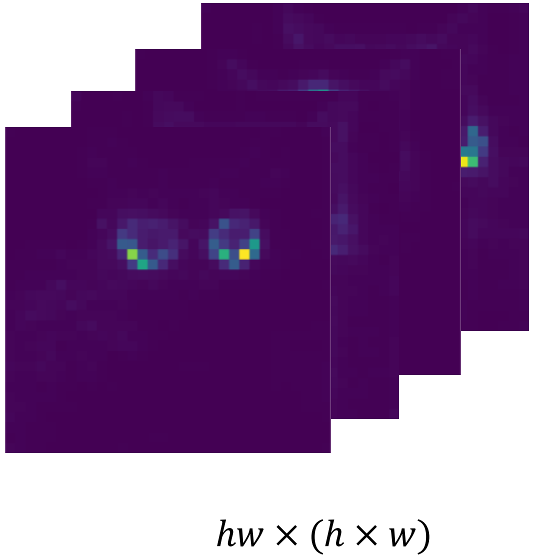
Self-Attention Maps

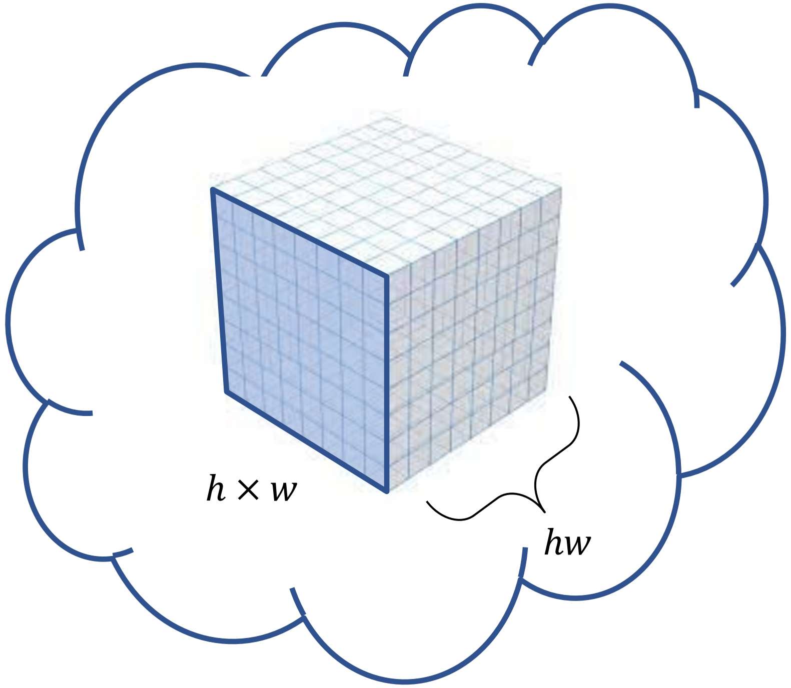


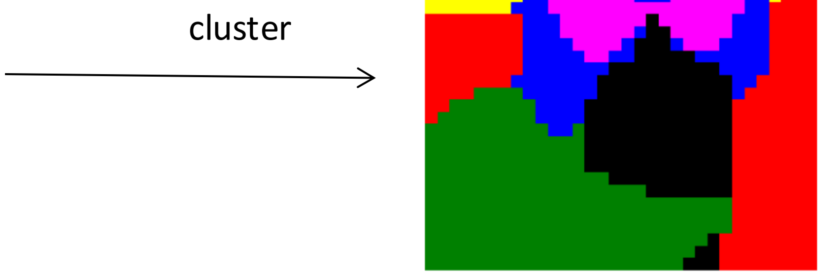
Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation [Tumanyan et al., CVPR 2023]

Self-Segmentation

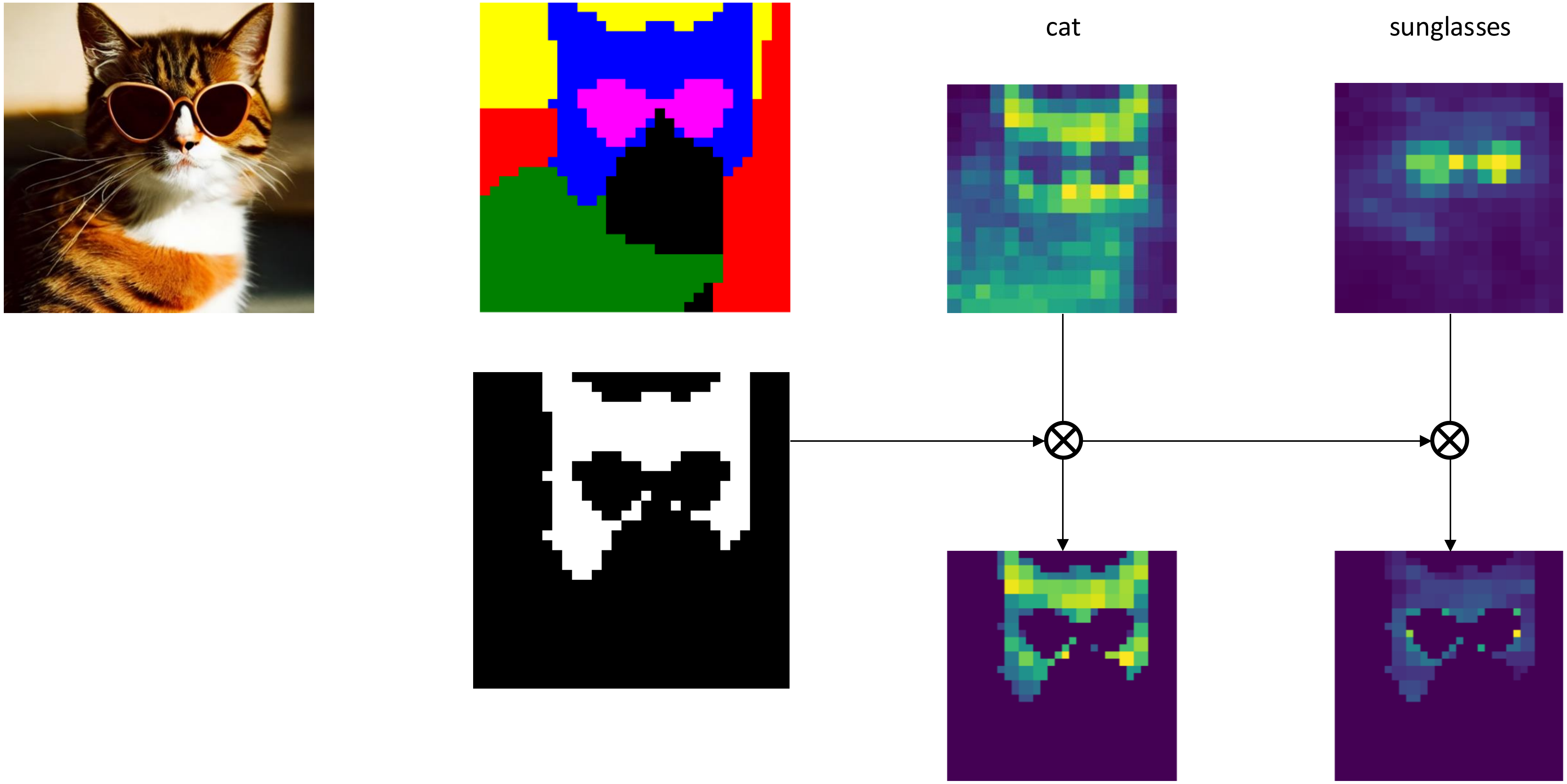
There is a lot of semantics in the self attention features!!!







Segments Labeling



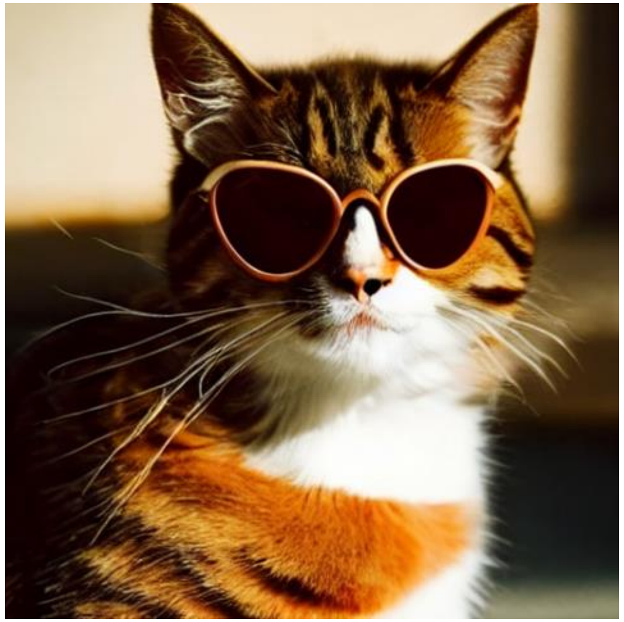
cat

sunglasses

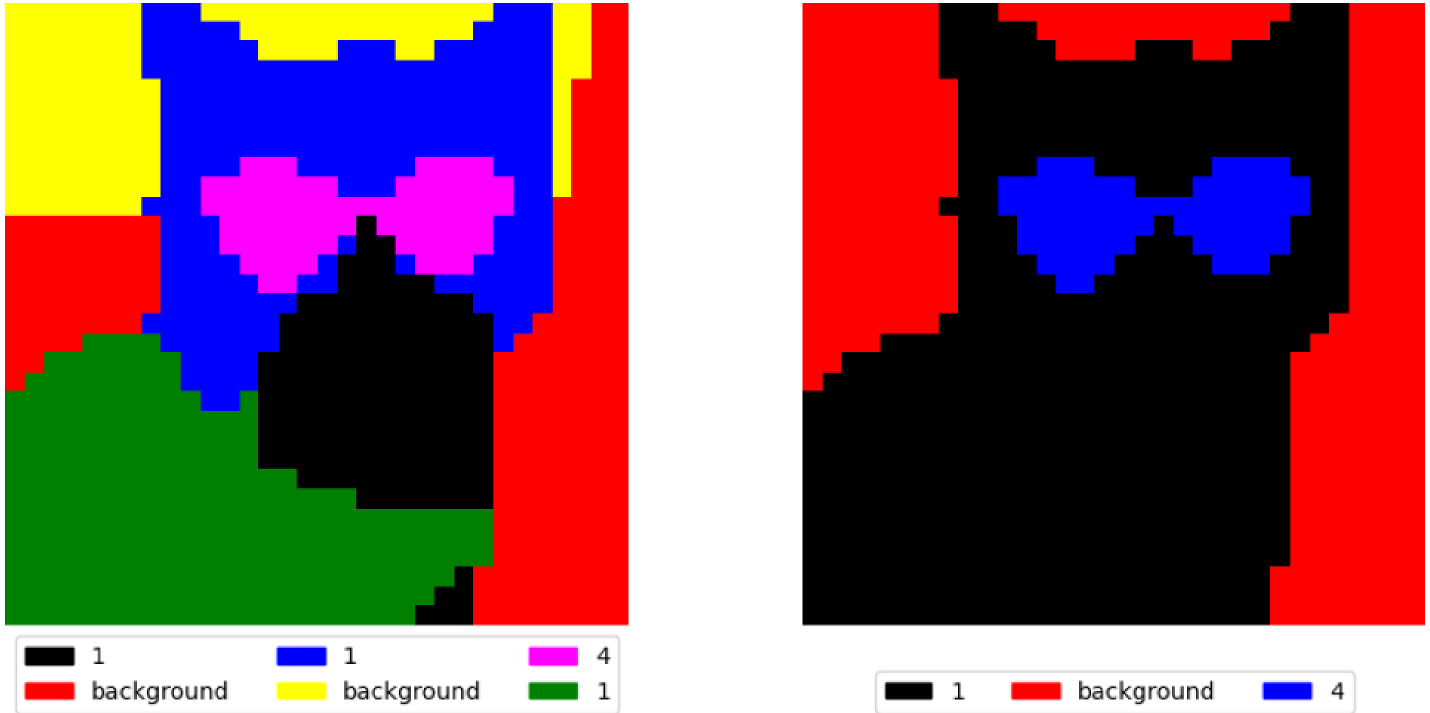
score: 0.65

score: 0.19

Segments Labling

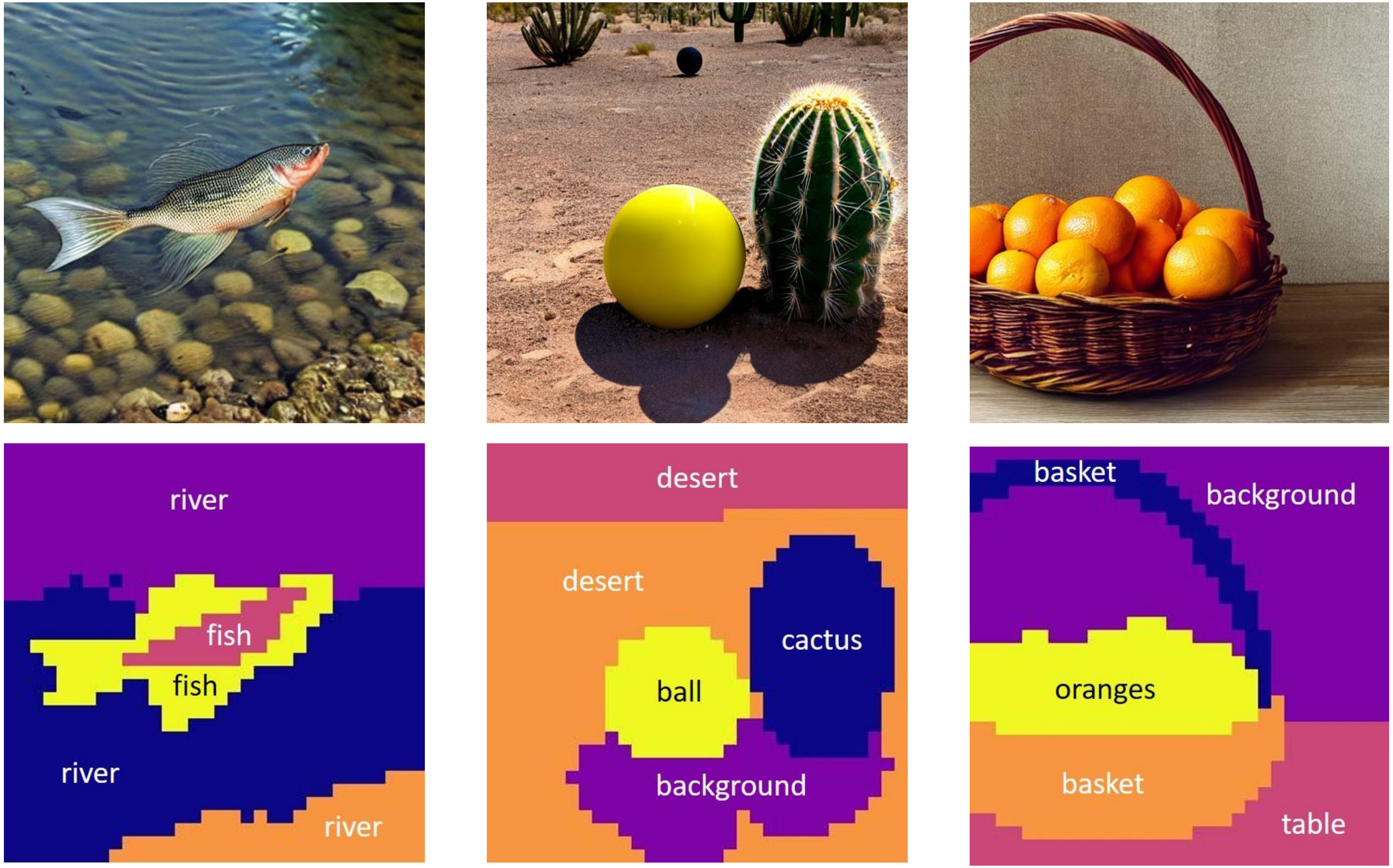


"a cat is wearing sunglasses"

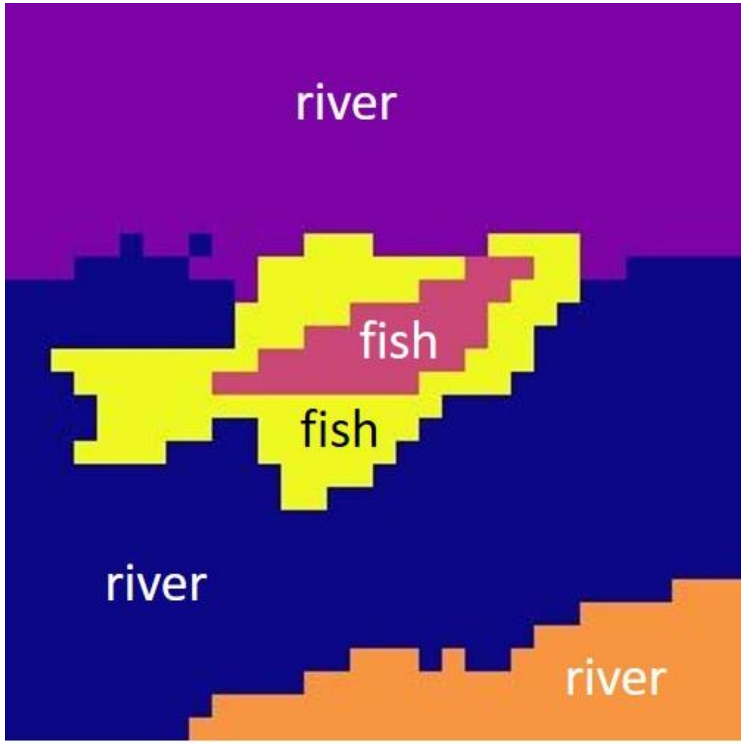


1-cat, 4-sunglasses

Self-Segmentation Results



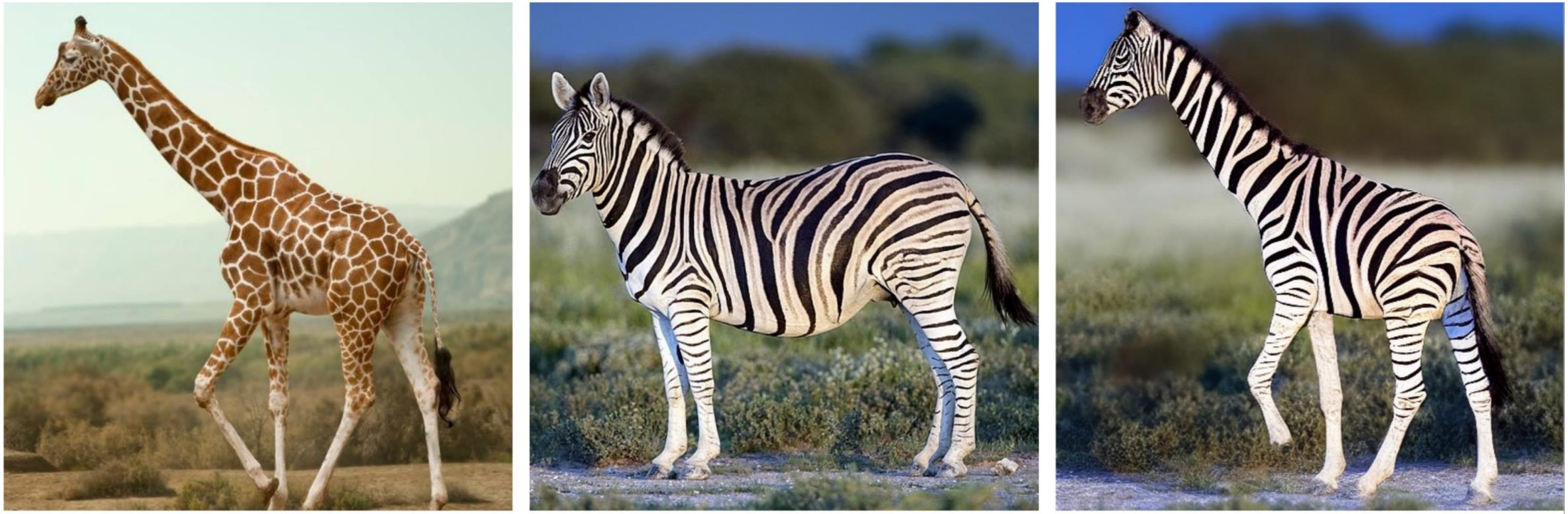
Self-Segmentation Results



Semantic Correspondence and Appearance Transfer

Cross-Image Attention [Alaluf, Y.*, Garibi, D.*, Patashnik, O., Averbuch-Elor, H., Cohen-Or, D., SIGGRAPH 2024]
DIFT [Tang et al., NeurIPS 2023]
A Tale of Two Features [Zhang et al., NeurIPS 2023]
Unsupervised Semantic Correspondence Using Stable Diffusion [Hedlin et al., NeurIPS 2023]
Diffusion Hyperfeatures [Luo et al., NeurIPS 2023]

Motivation

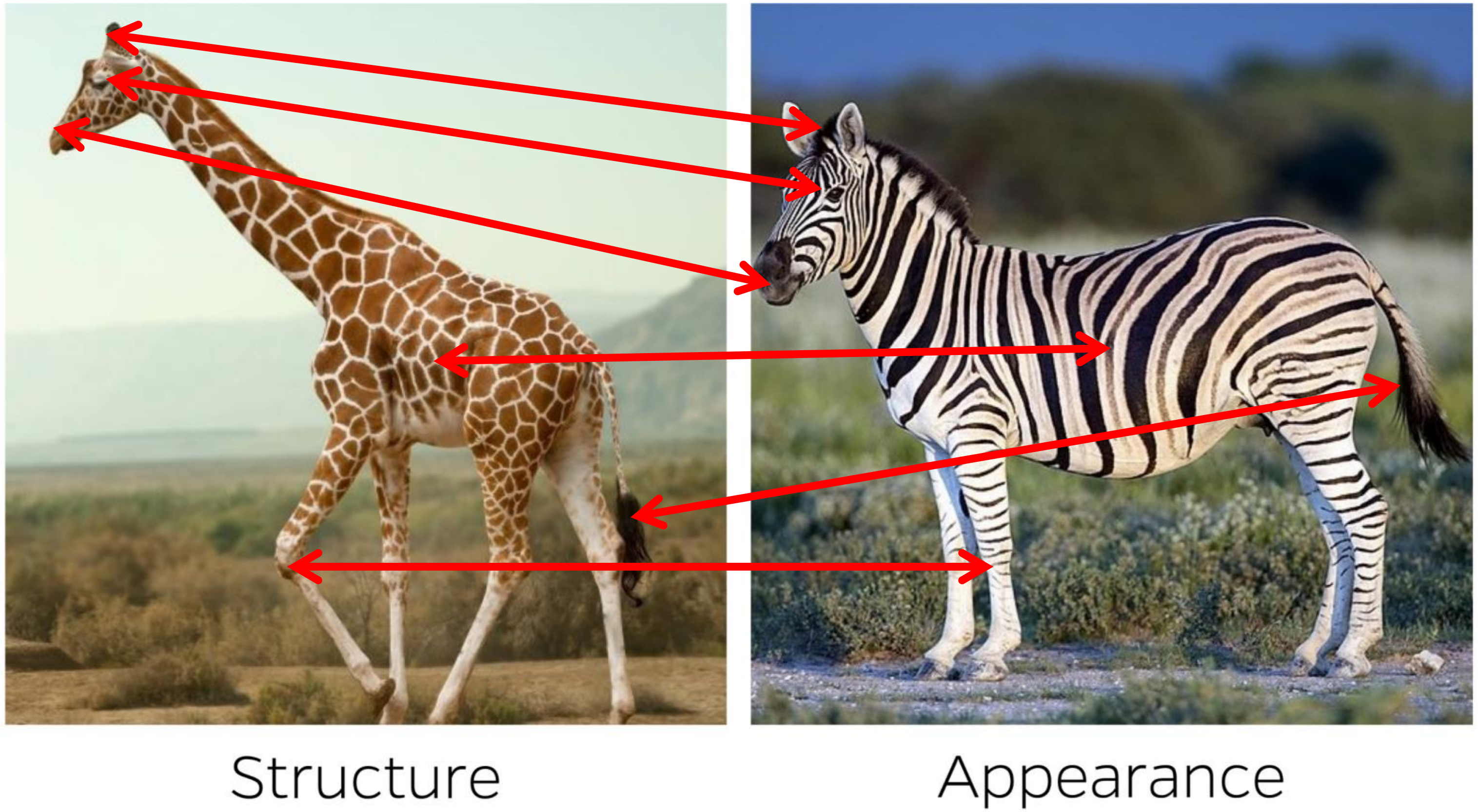


Structure

Appearance

Output

Motivation

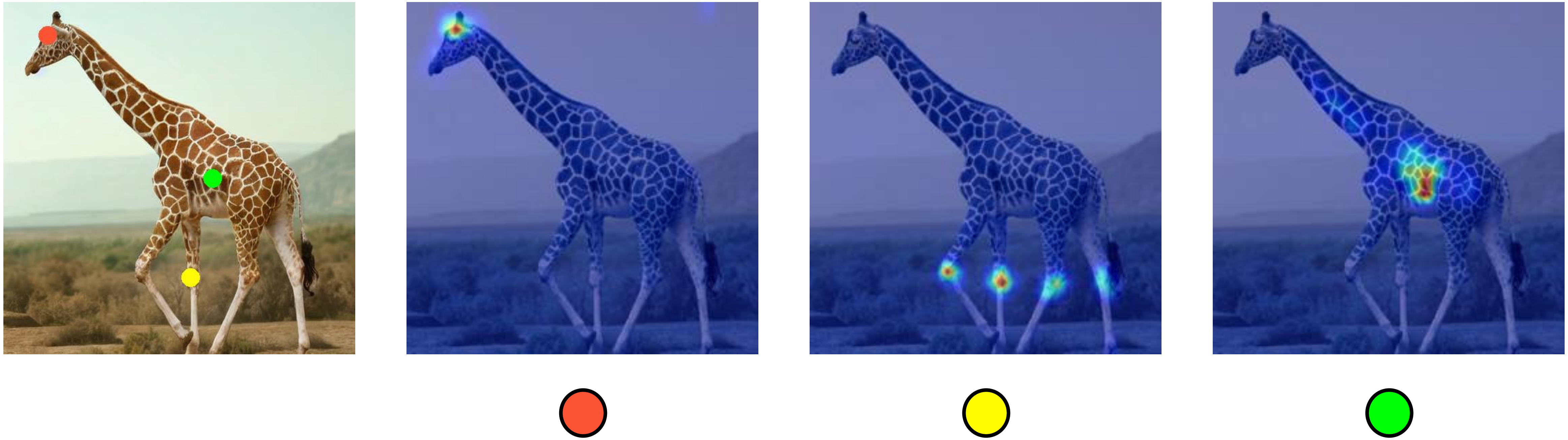


Main challenge is to find semantic correspondences between the images

Attention Is All You Need

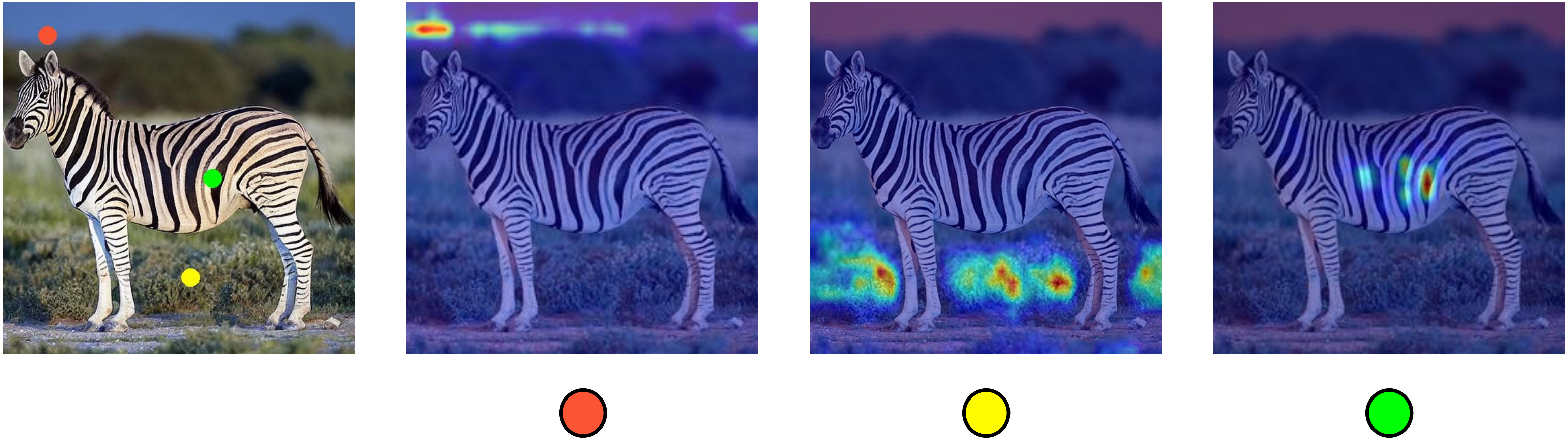
Attention Is All You Need

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$



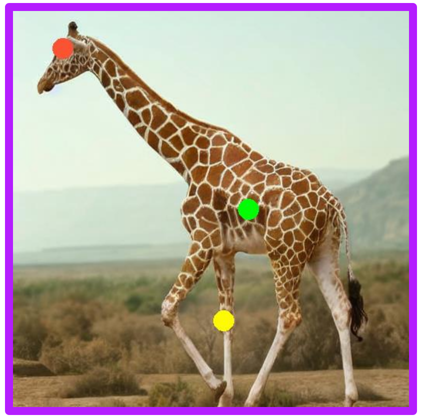
Attention Is All You Need

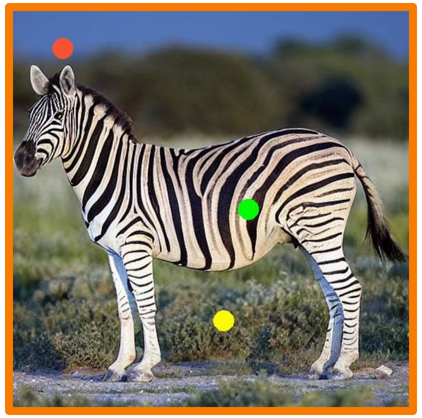
$$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$



Attention Is All You Need

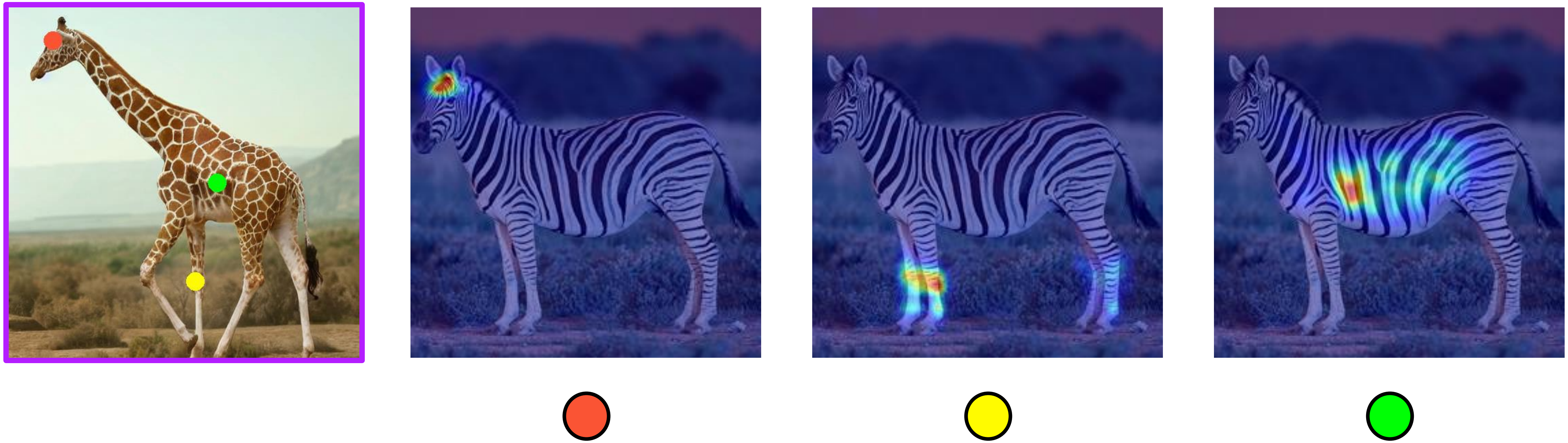
$$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$





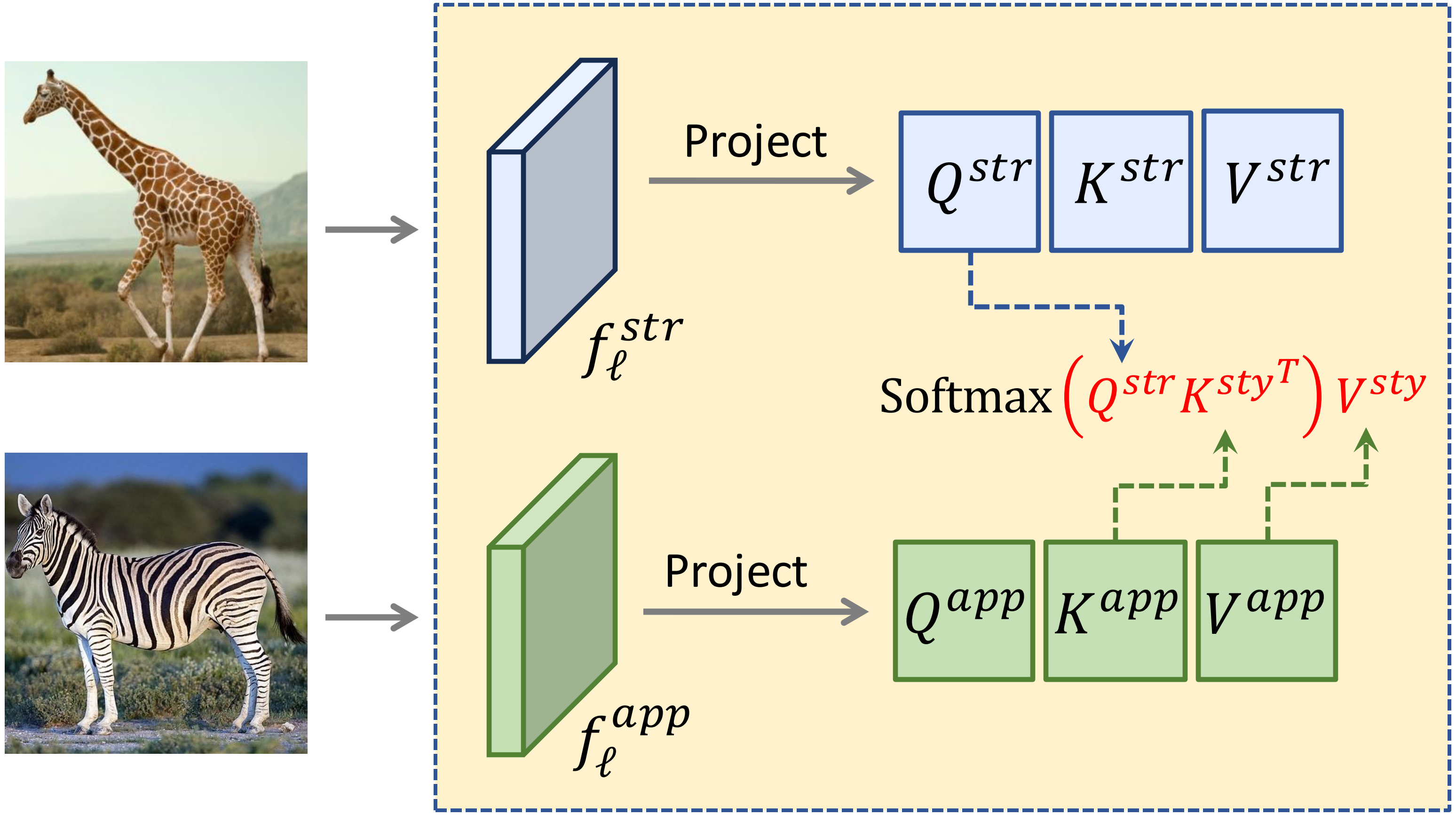
Attention Is All You Need

$$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

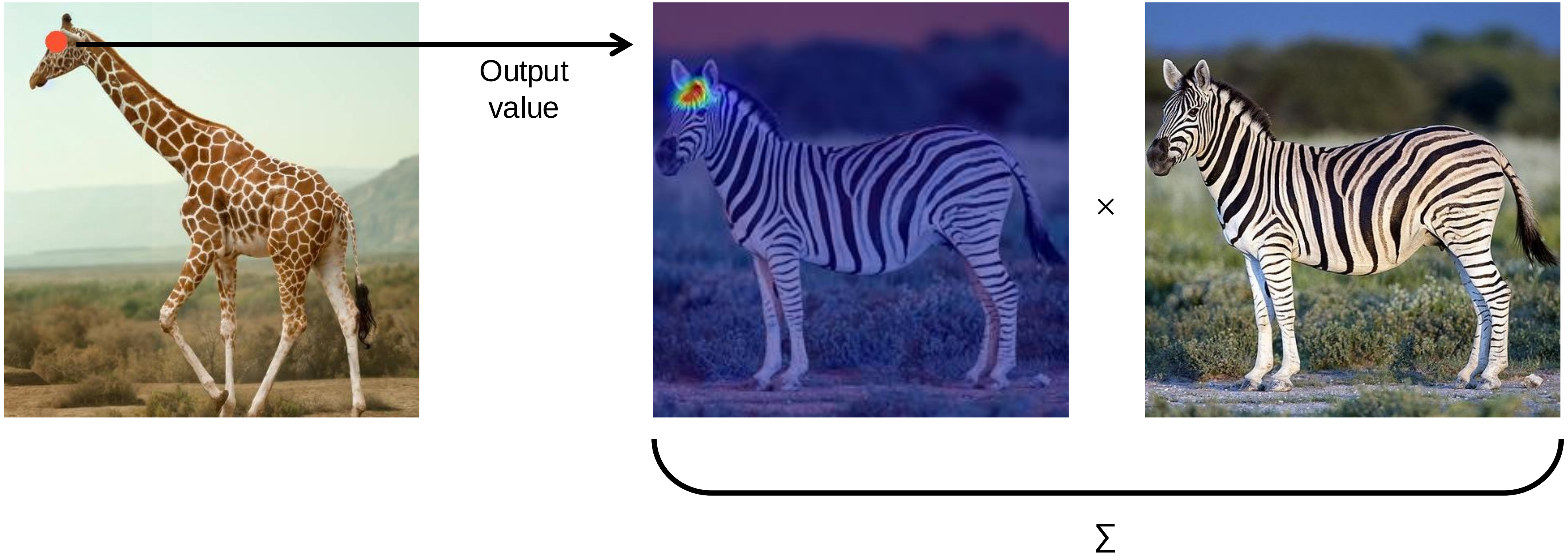


Cross-Image Attention

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



Cross-Image Attention

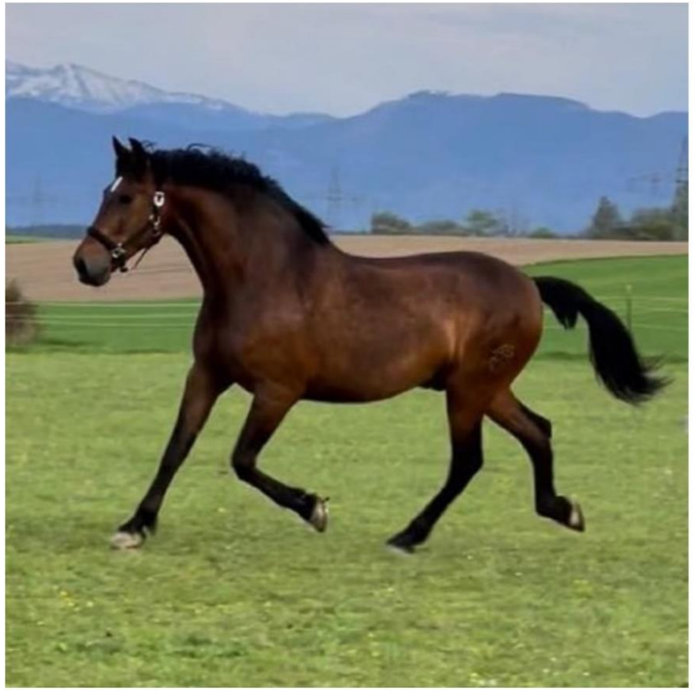


Output value

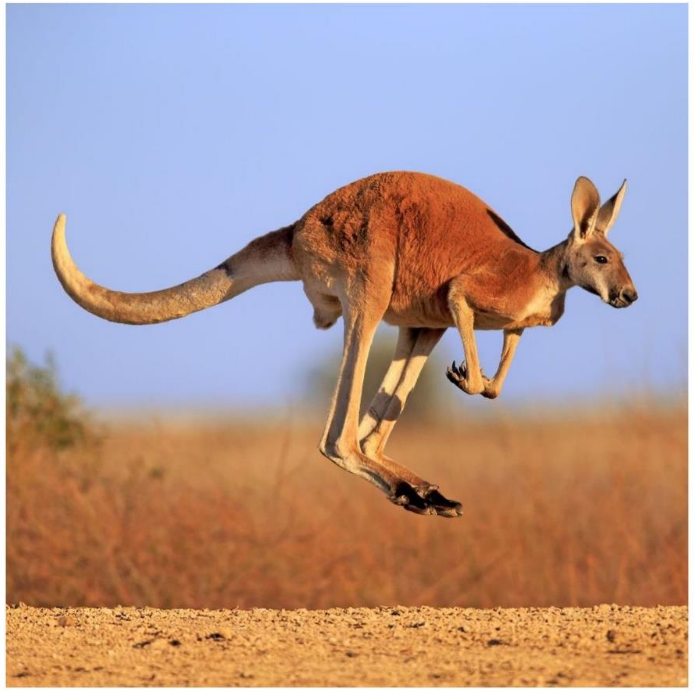
×

Σ

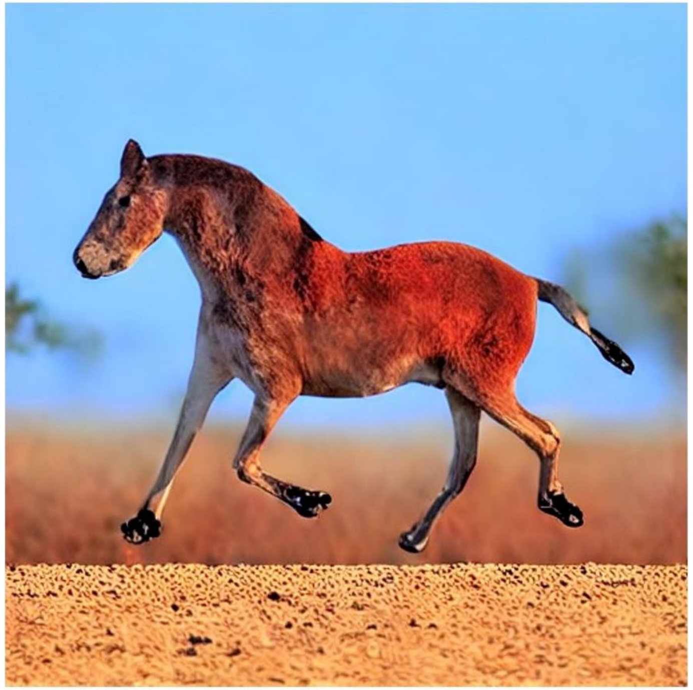
Appearance Transfer Results



Structure

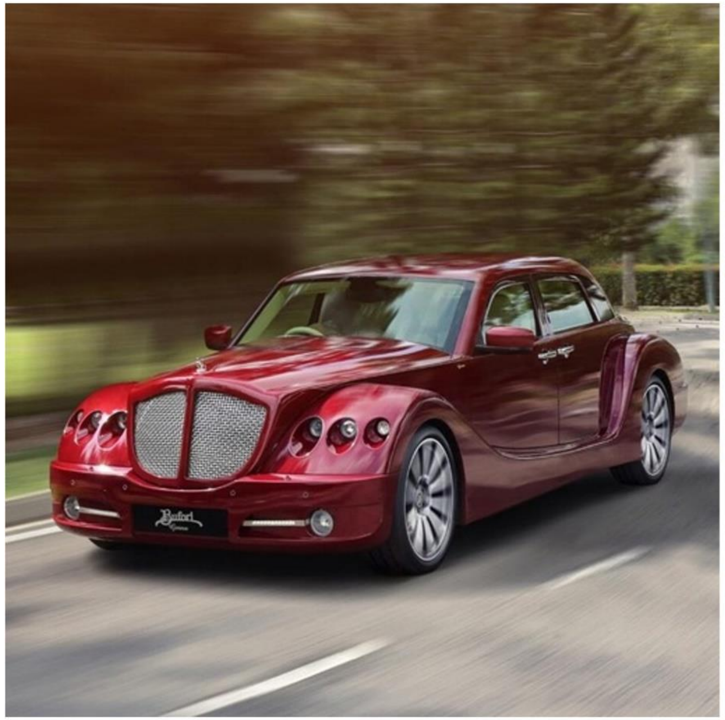


Appearance

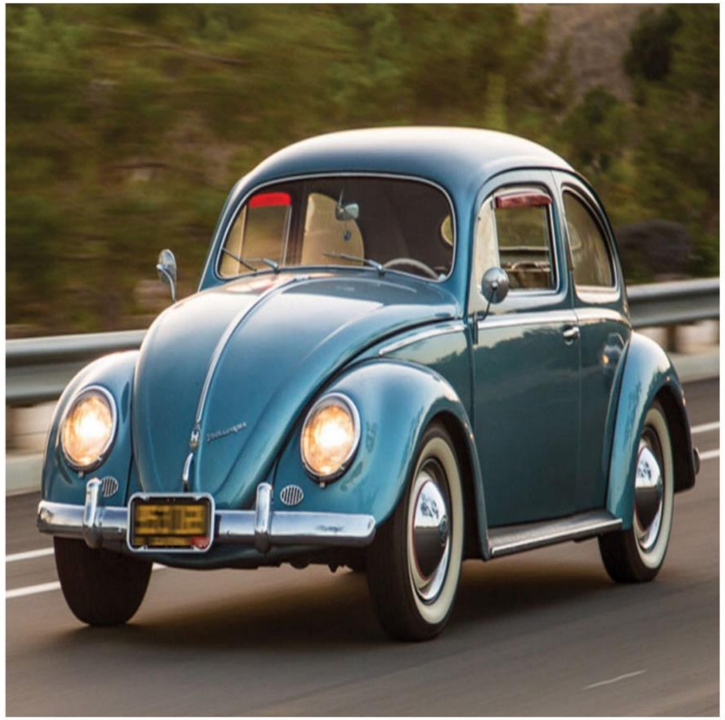


Output

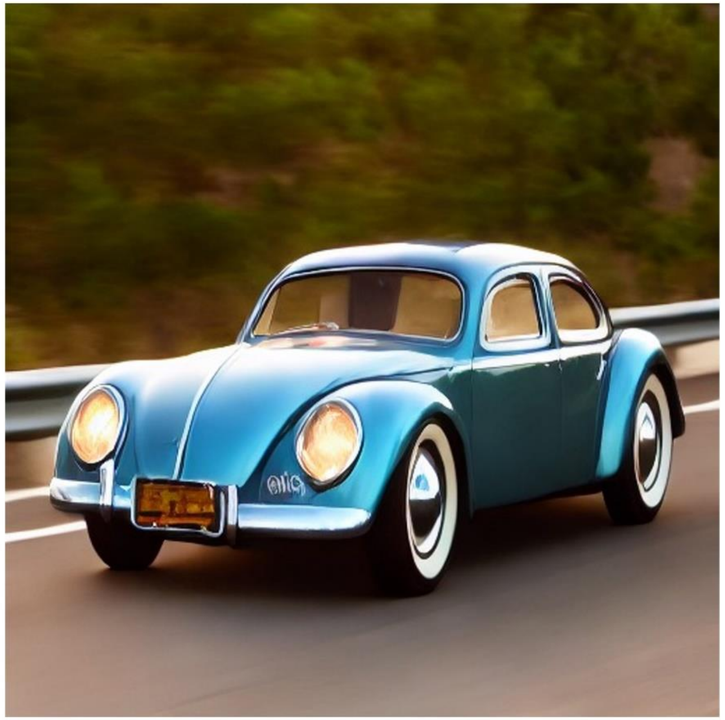
Appearance Transfer Results



Structure

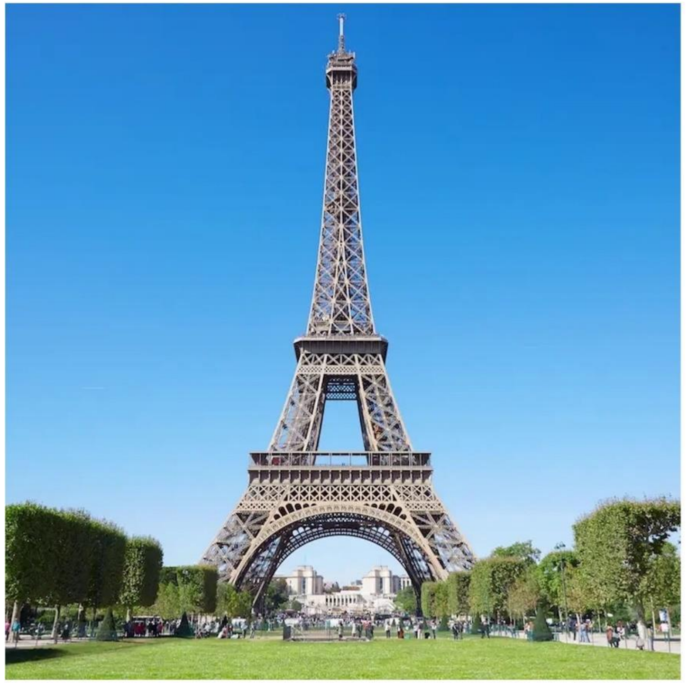


Appearance

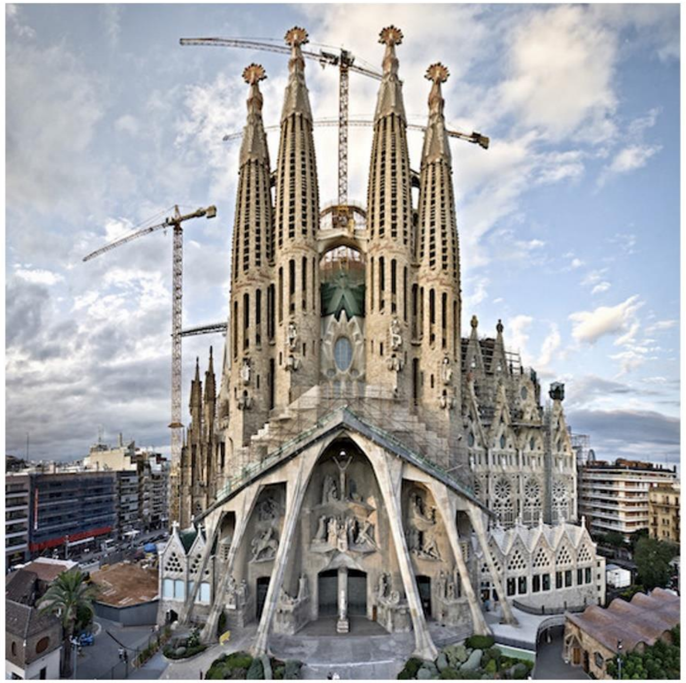


Output

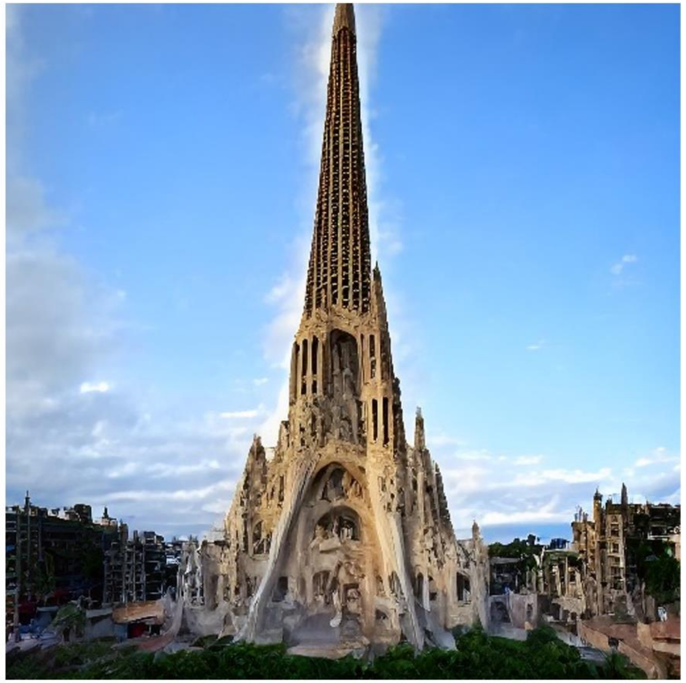
Appearance Transfer Results



Structure

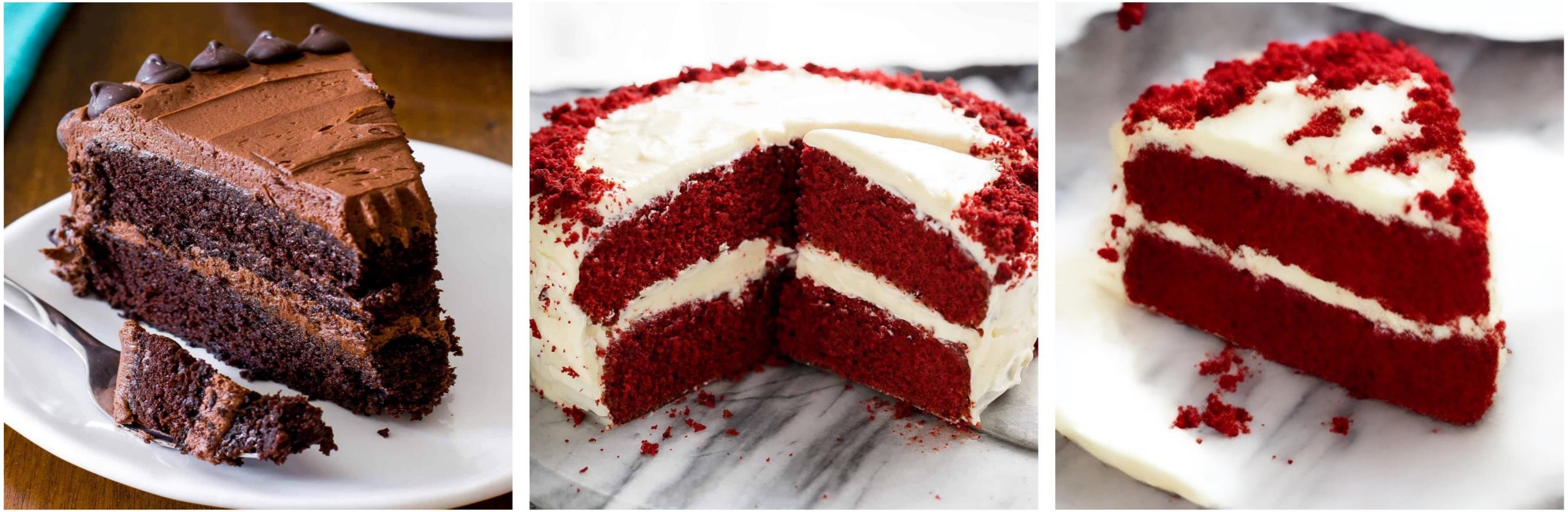


Appearance



Output

Appearance Transfer Results

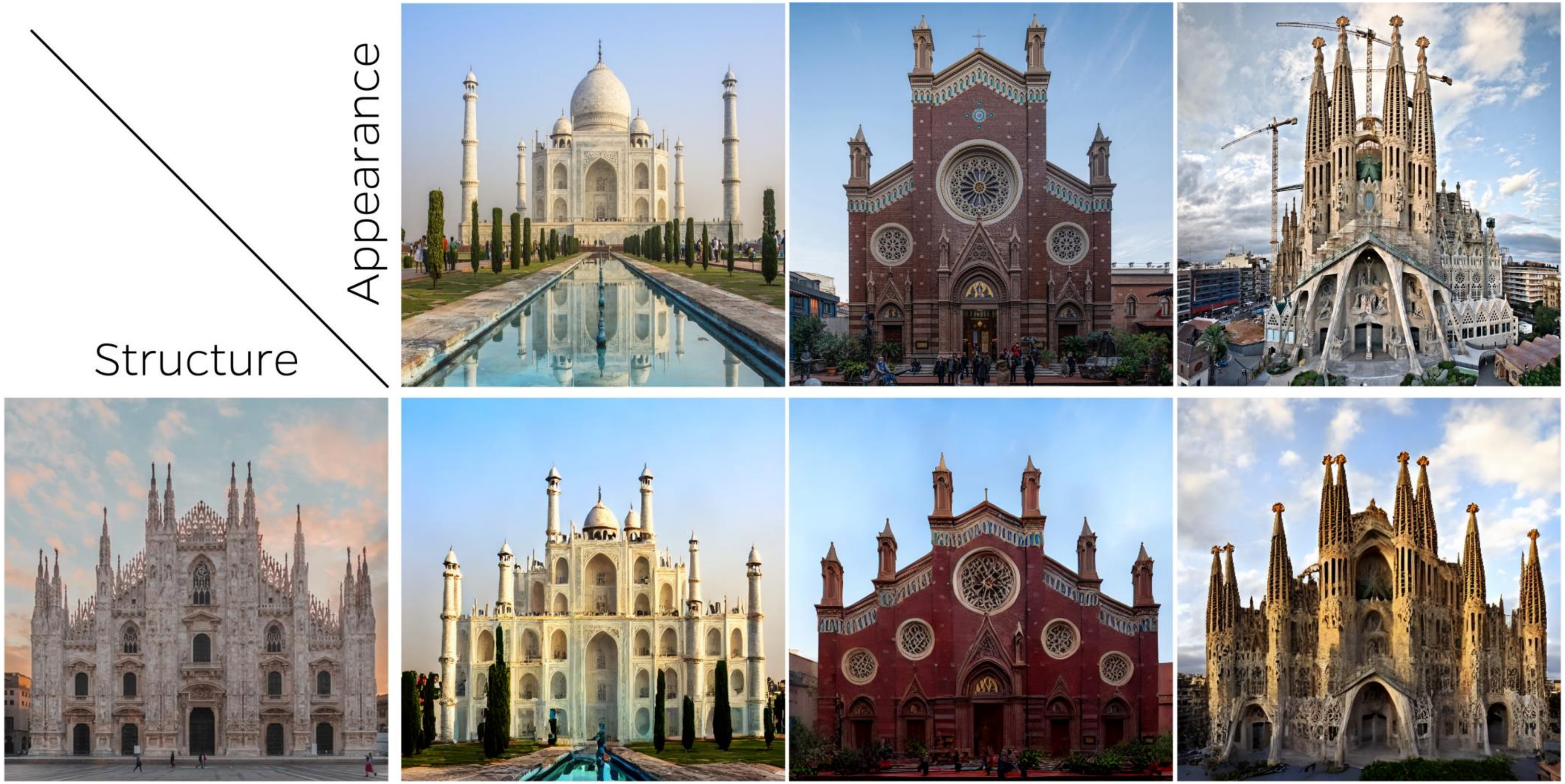


Structure

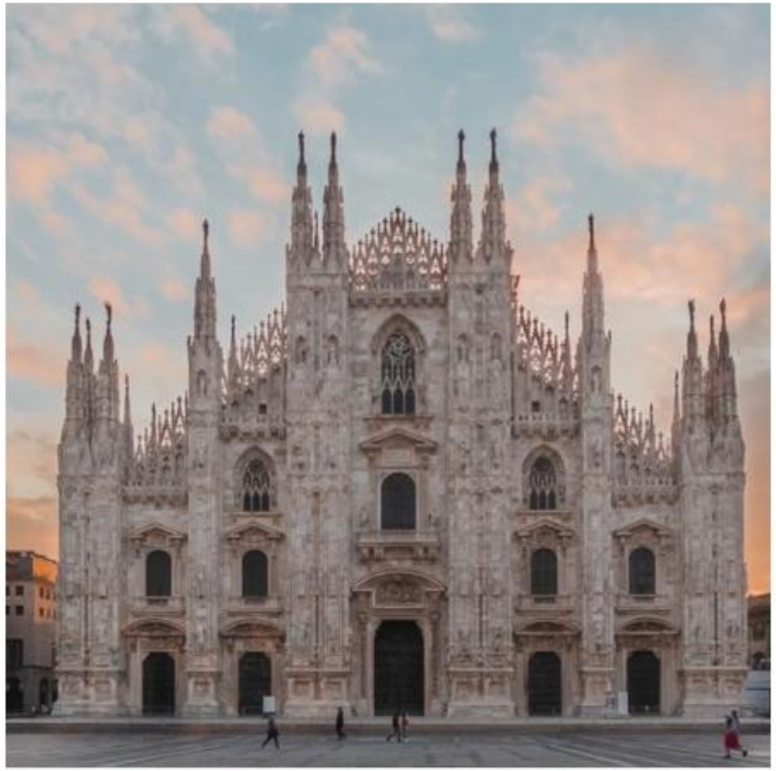
Appearance

Output

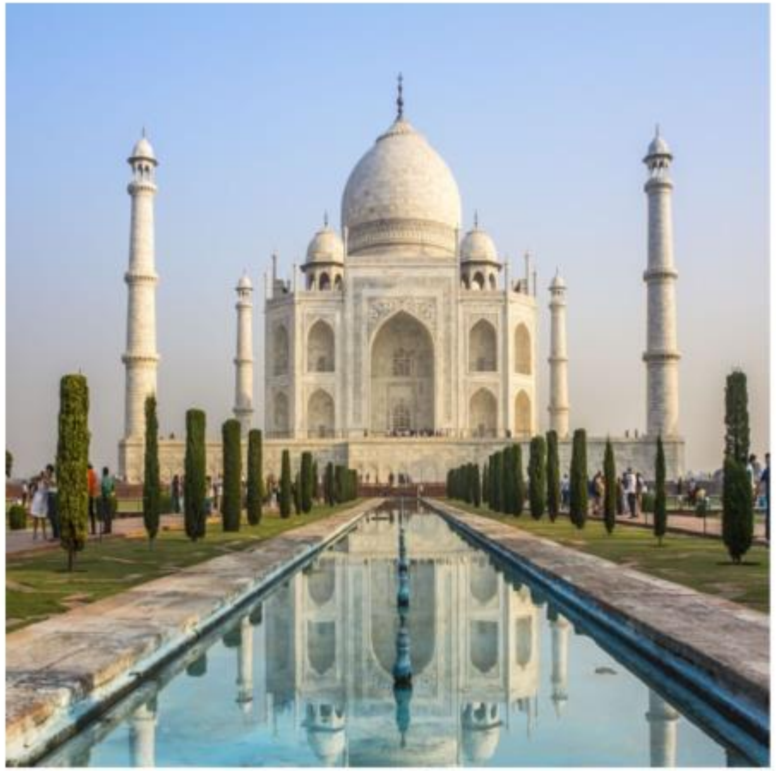
Appearance Transfer Results



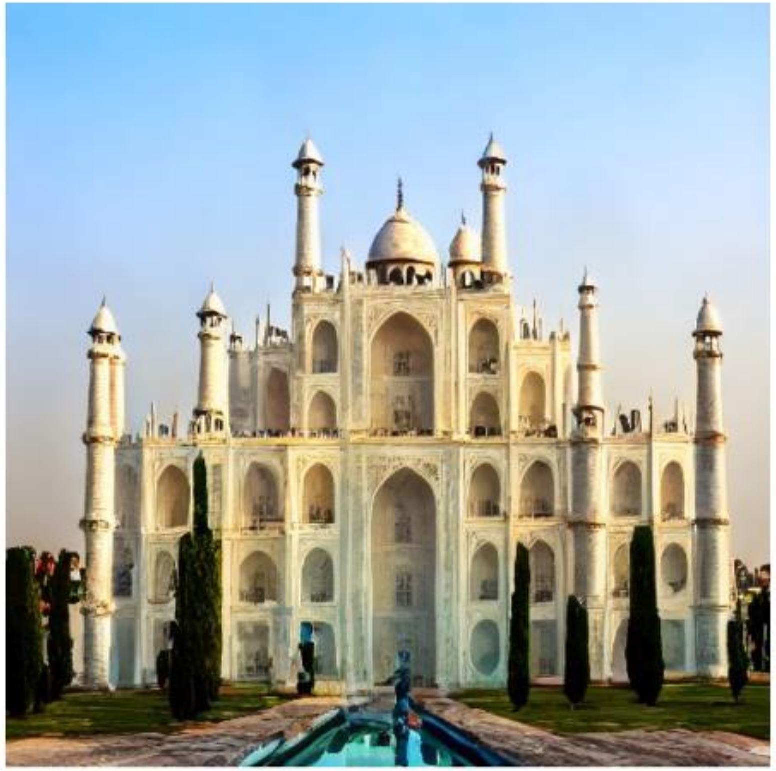
Appearance Transfer Results



Structure



Appearance

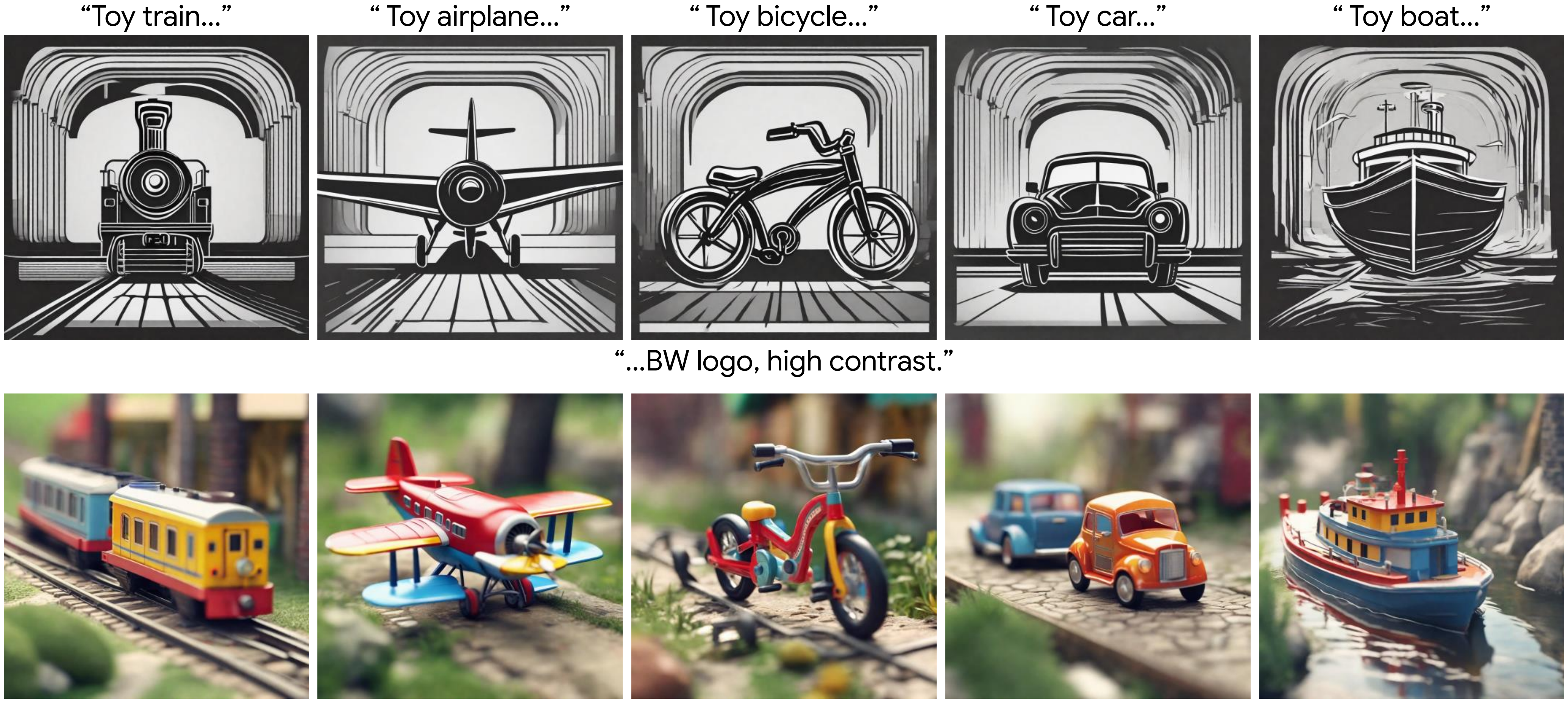


Output

Consistent Generation

Style Aligned Image Generation via Shared Attention [Hertz, A.*, Voynov, A.*, Fruchter, S., Cohen-Or, D. CVPR 2024]
Tune-A-Video [Wu et al., ICCV 2023]
Pix2Video [Ceylan et al., ICCV 2023]
Text2Video-Zero [Khachatryan et al., ICCV 2023]
TokenFlow [Geyer et al., ICLR 2024]
ConsiStory [Tewel el al., SIGGRAPH 2024]
AnimateAnyone [Hu et al., 2023]
MagicAnimate [Xu et al., 2023]

Style Aligned



“Toy train…”

“ Toy airplane…”

“ Toy bicycle…”

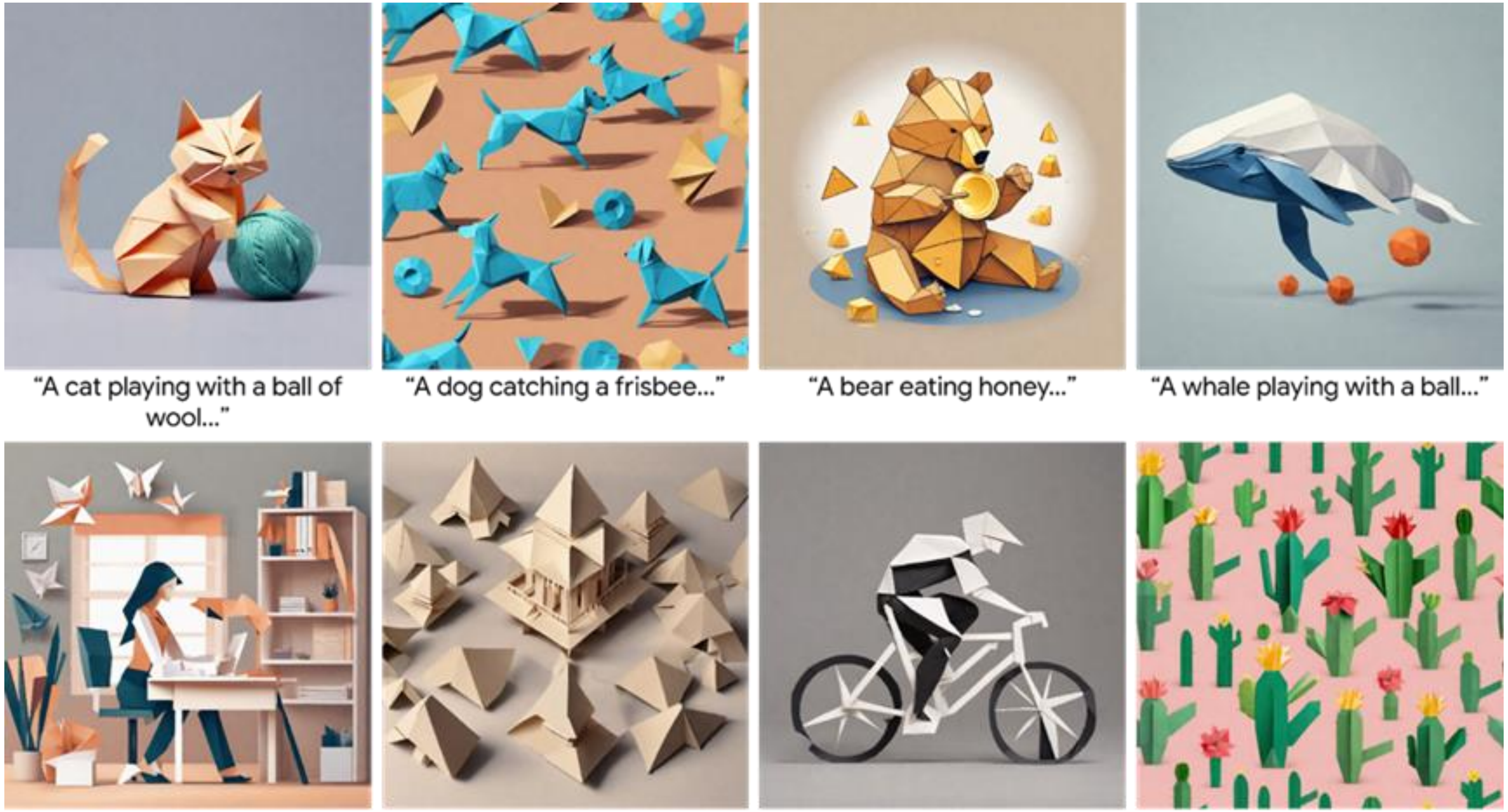
“ Toy car…”

“ Toy boat…”

“…BW logo, high contrast.”

“…colorful, macro photo.”

Text-to-Image Generation



"A cat playing with a ball of wool..."

"A dog catching a frisbee..."

"A bear eating honey..."

"A whale playing with a ball..."

"A woman working in the office..."

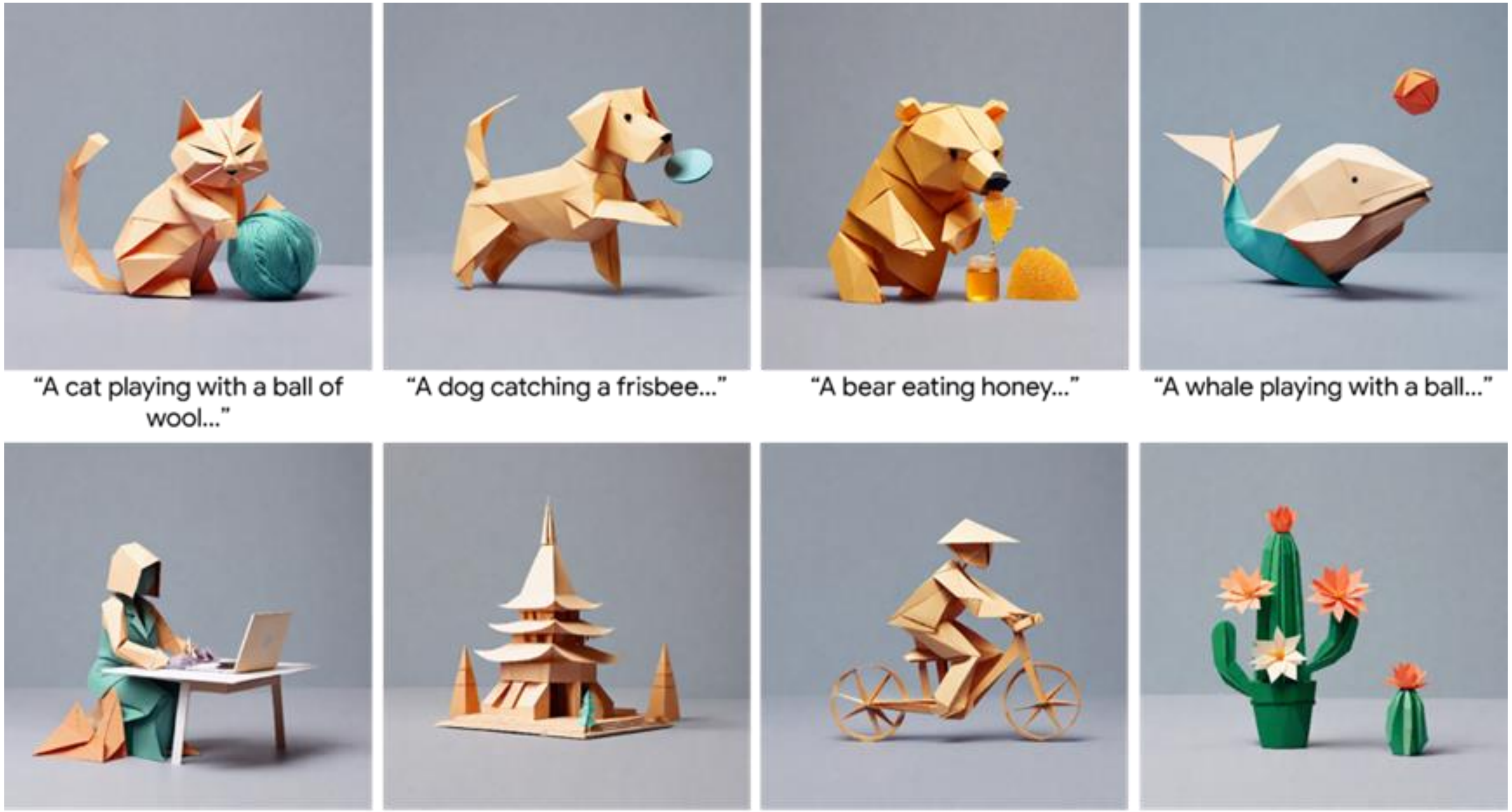
"A temple..."

"A person riding a bike..."

"A cactus..."

"... in minimal origami style."

Text-to-Image Generation with Style Aligned



"A cat playing with a ball of wool…"

"A dog catching a frisbee…"

"A bear eating honey…"

"A whale playing with a ball…"

"A woman working in the office…"

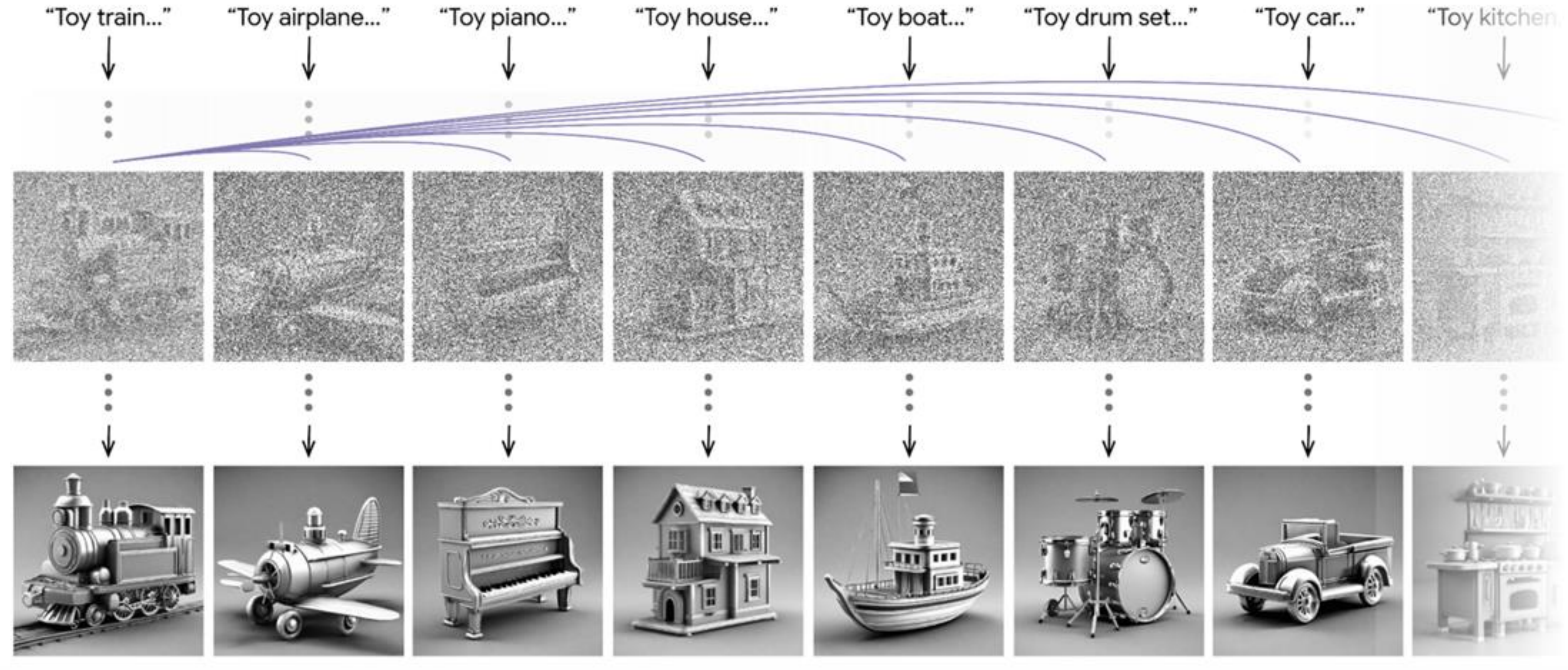
"A temple…"

"A person riding a bike…"

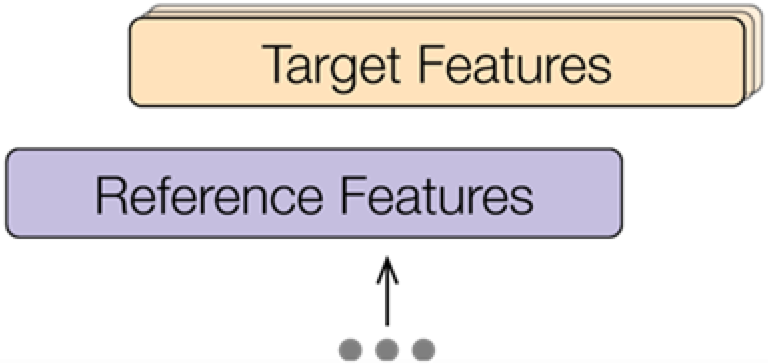
"A cactus…"

"… in minimal origami style."

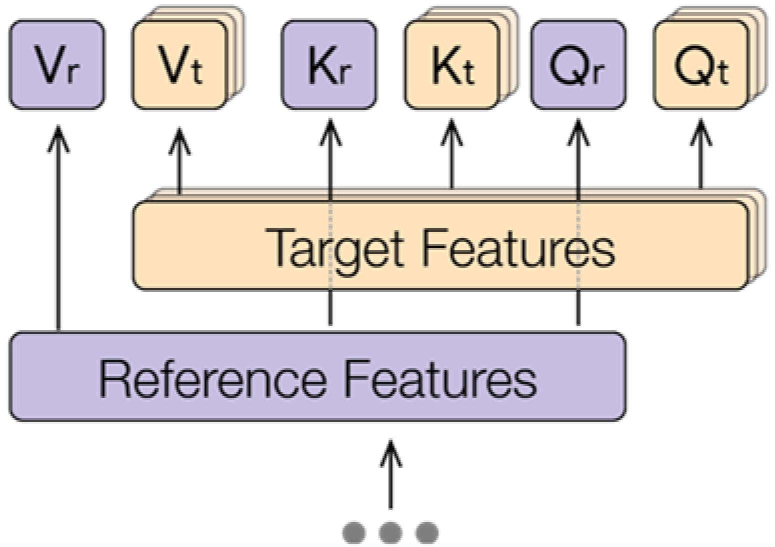
Shared Attention During the Diffusion Process



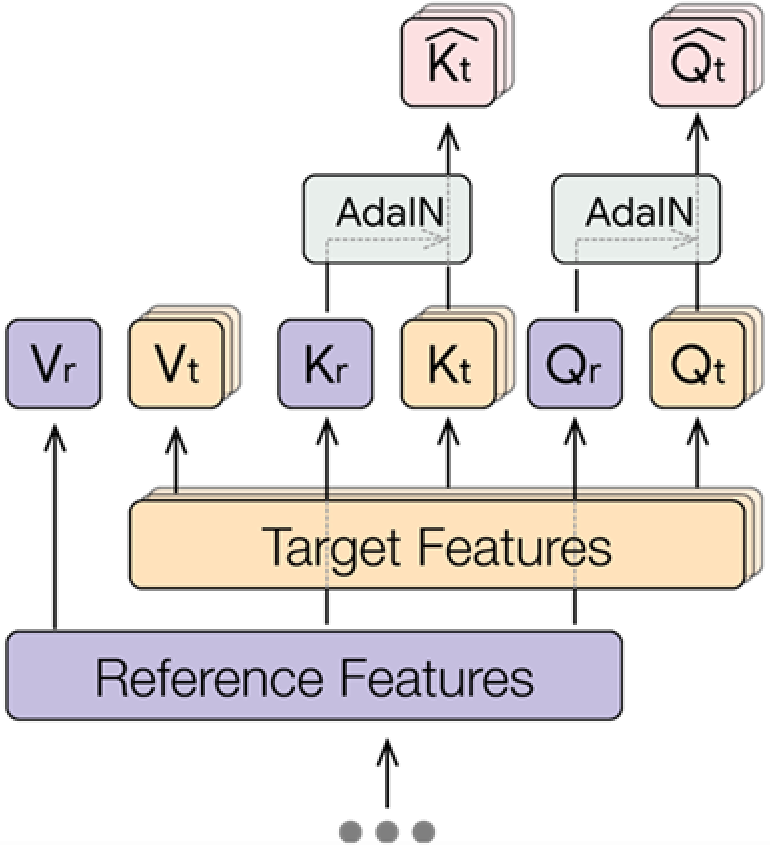
Shared Attention Layer



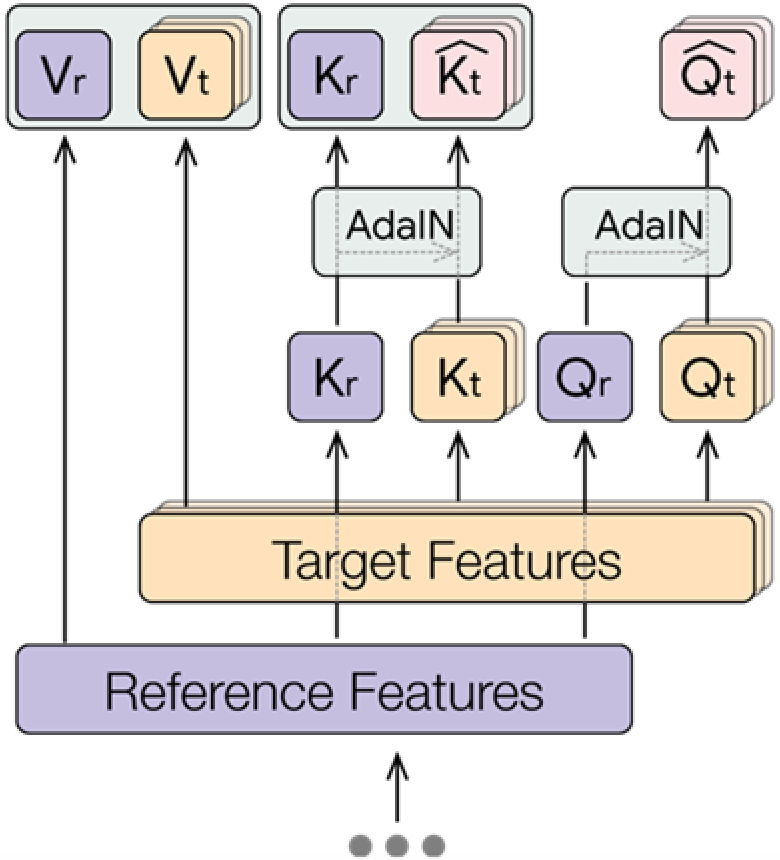
Shared Attention Layer



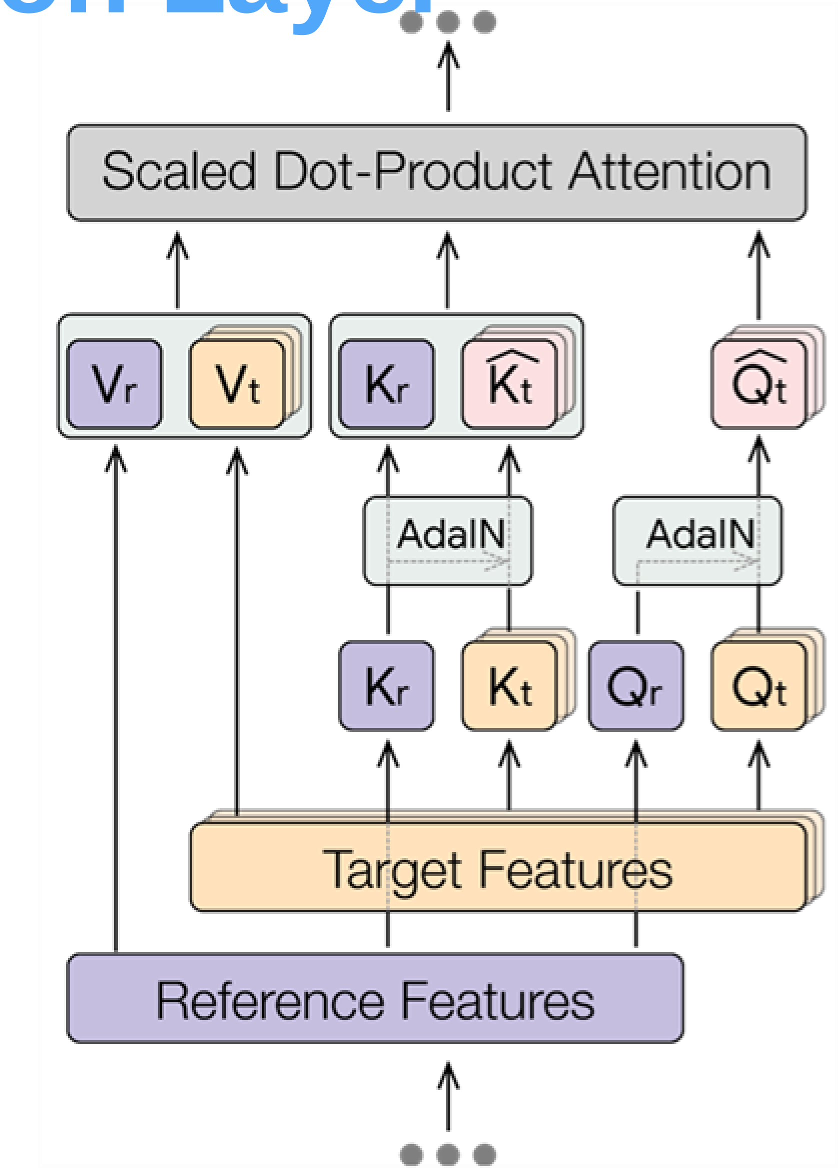
Shared Attention Layer



Shared Attention Layer

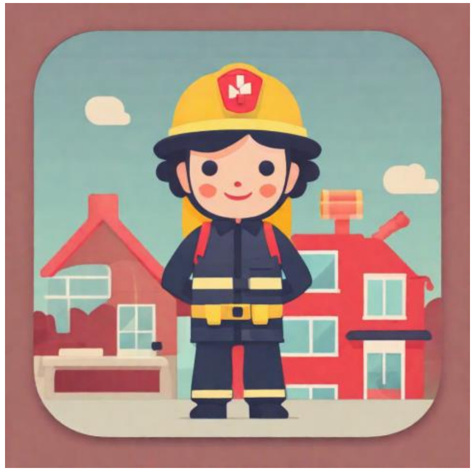


Shared Attention Layer

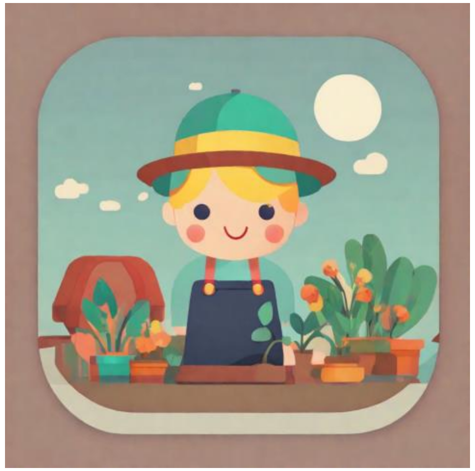


Style Aligned Generation of Synthetis Images

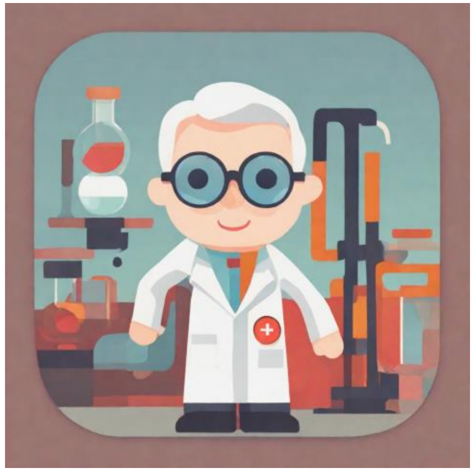
“Firewoman…”



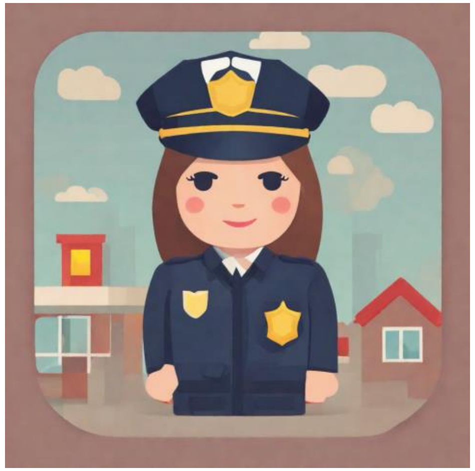
“ Gardner…”



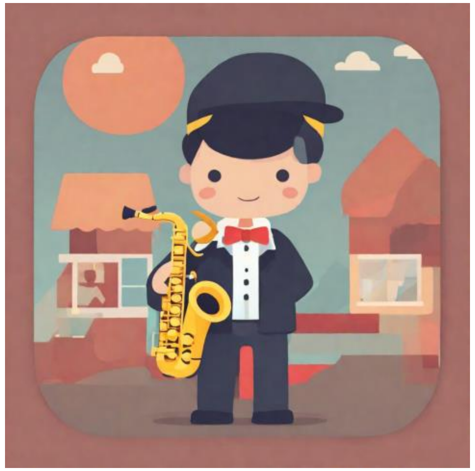
“ Scientist …”



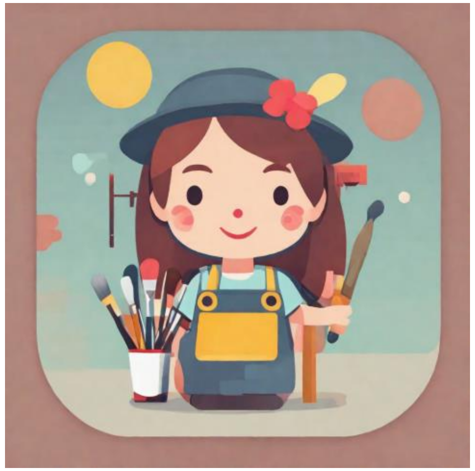
“ Police woman…”



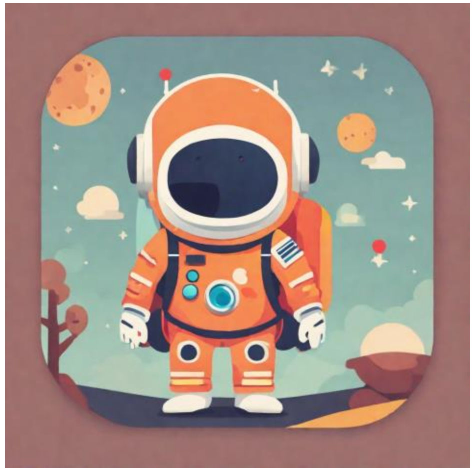
“ Saxophone player…”



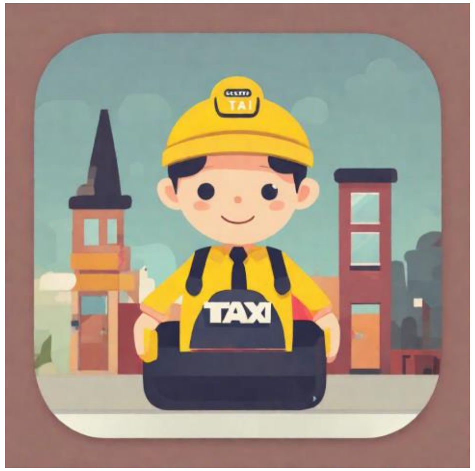
“Painter…”



“Astronaut…”

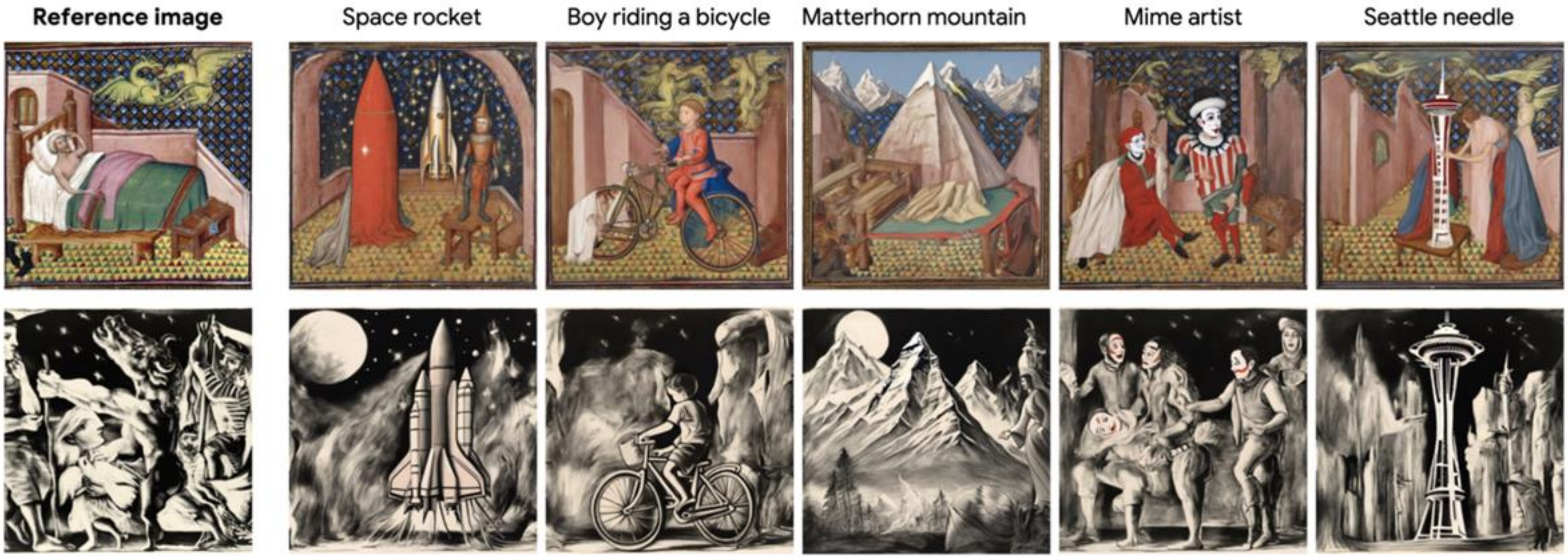


“Taxi Driver…”

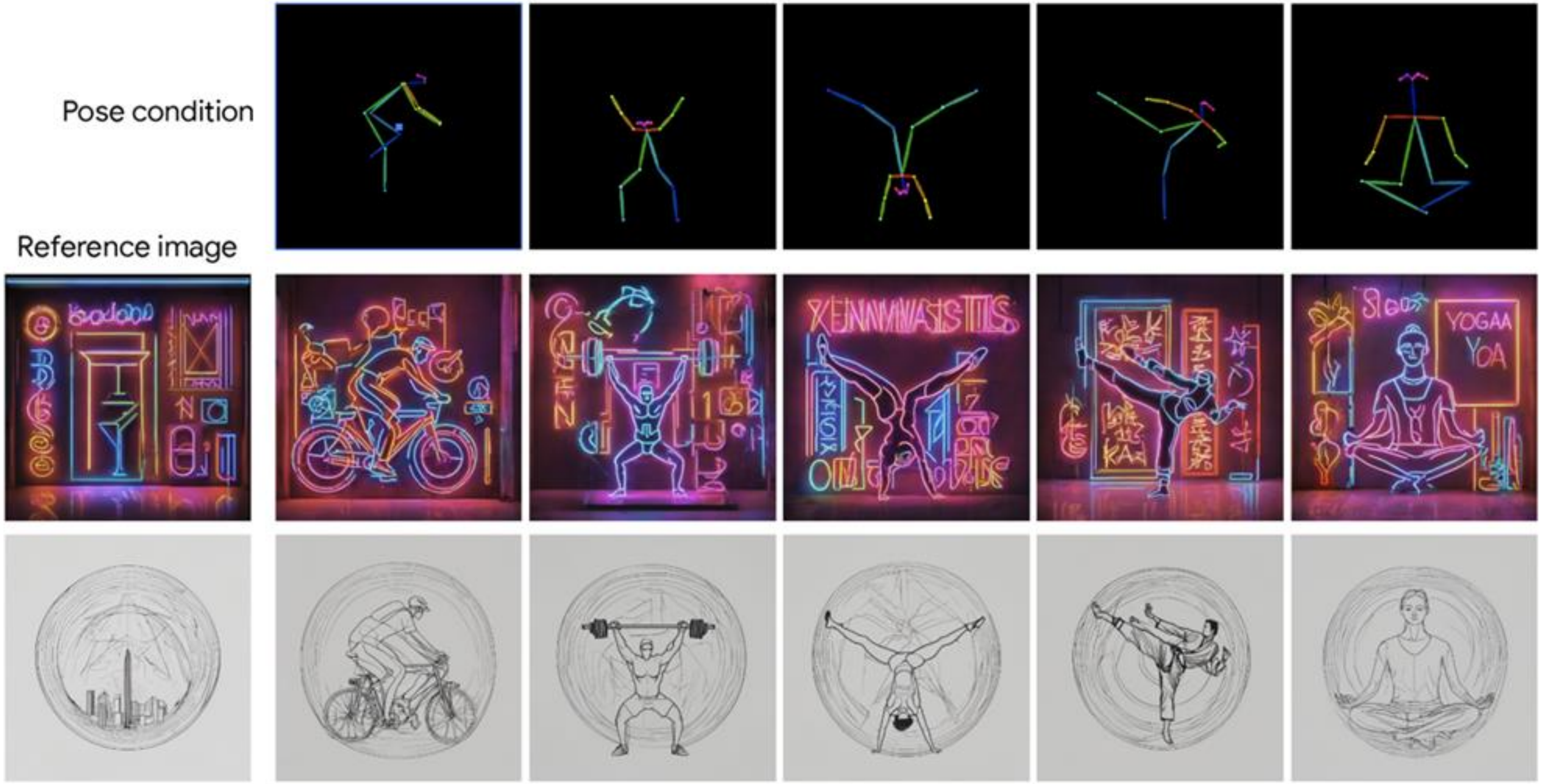


“ …in minimal flat design illustartion.”

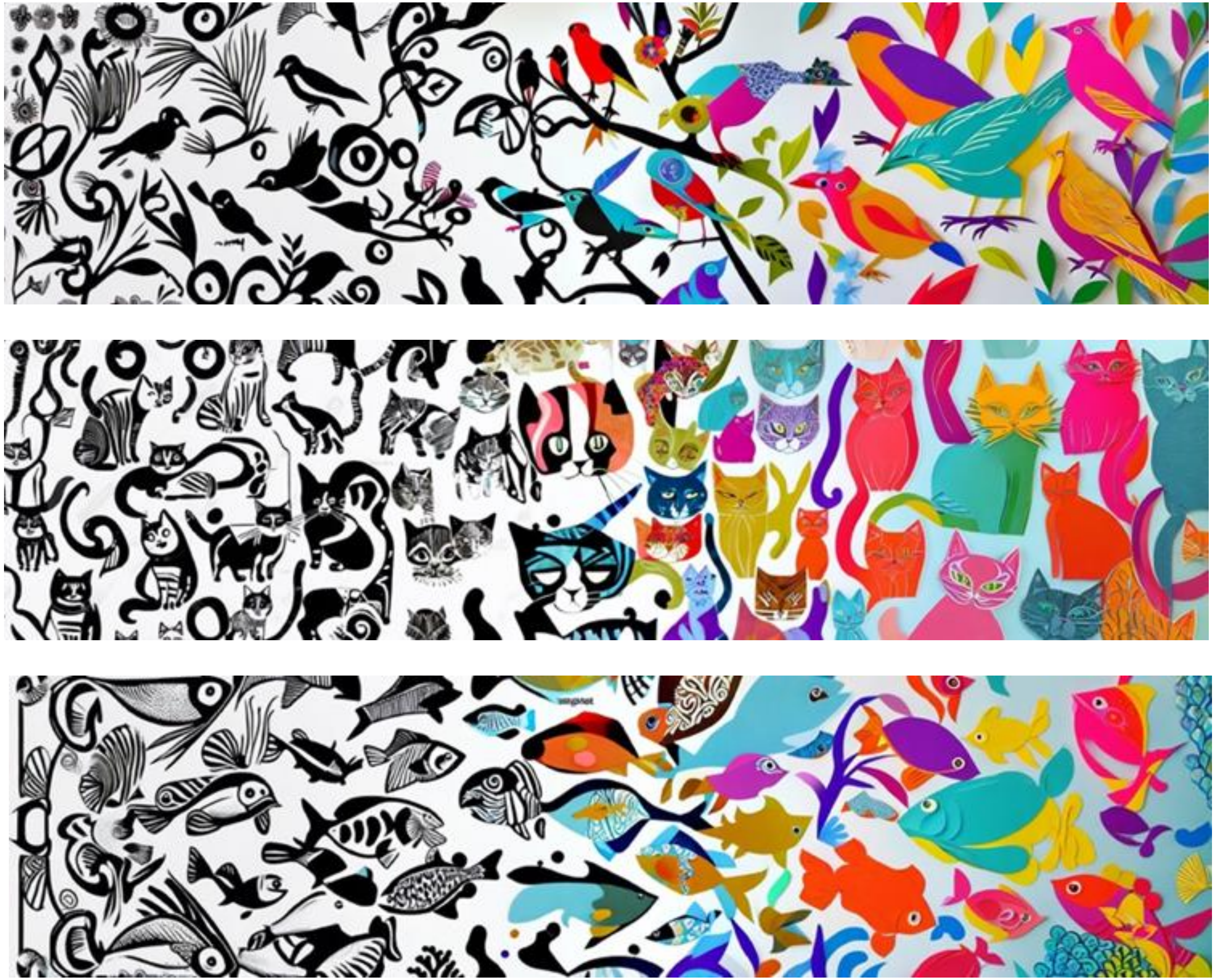
Style Aligned Generation from an Input Image



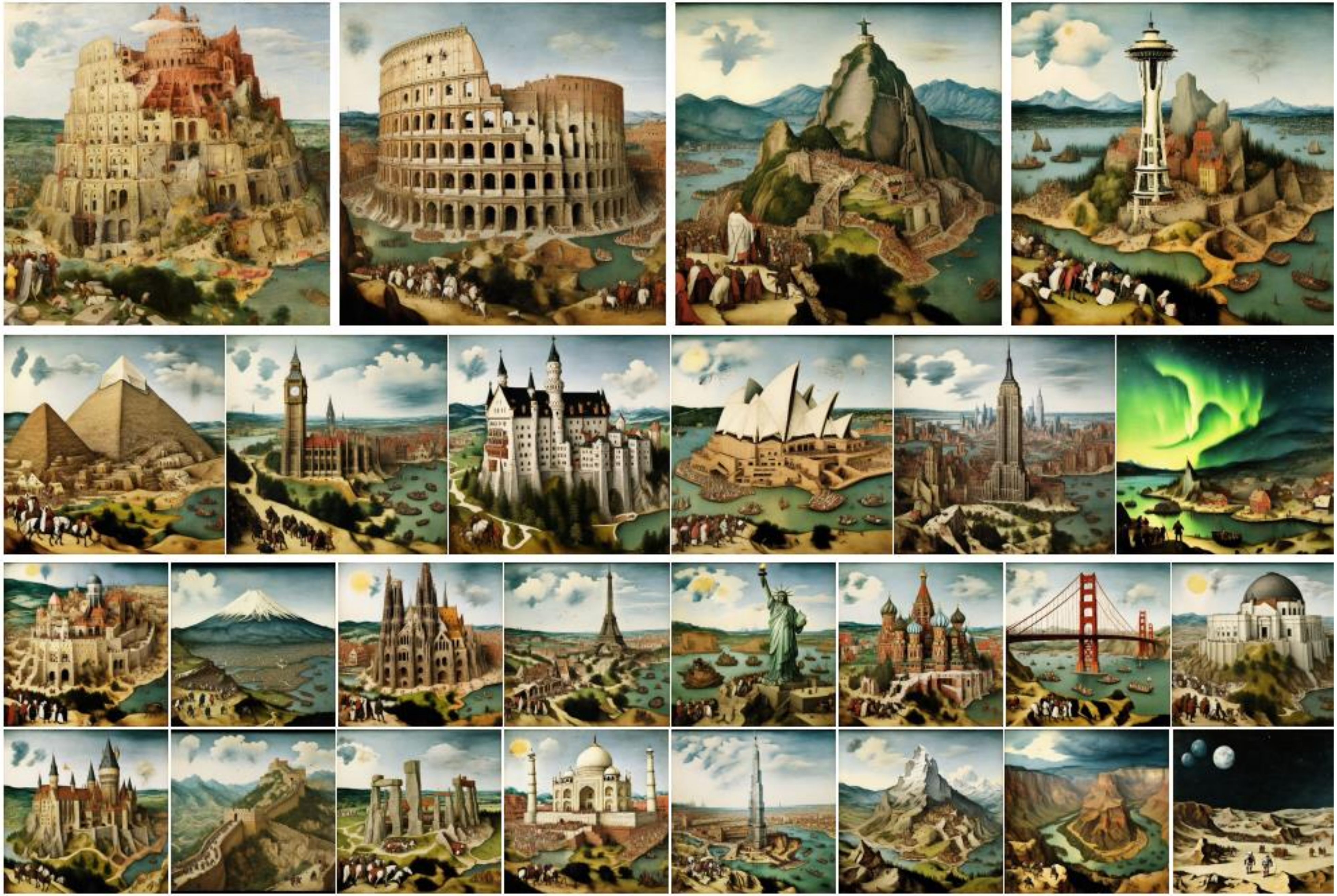
ControlNet + Style Aligned



Left Reference



Right Reference



Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

Break

Personalization and Editing

Beyond Single (RGB) Image Generation

Diffusion Models for 3D Generation