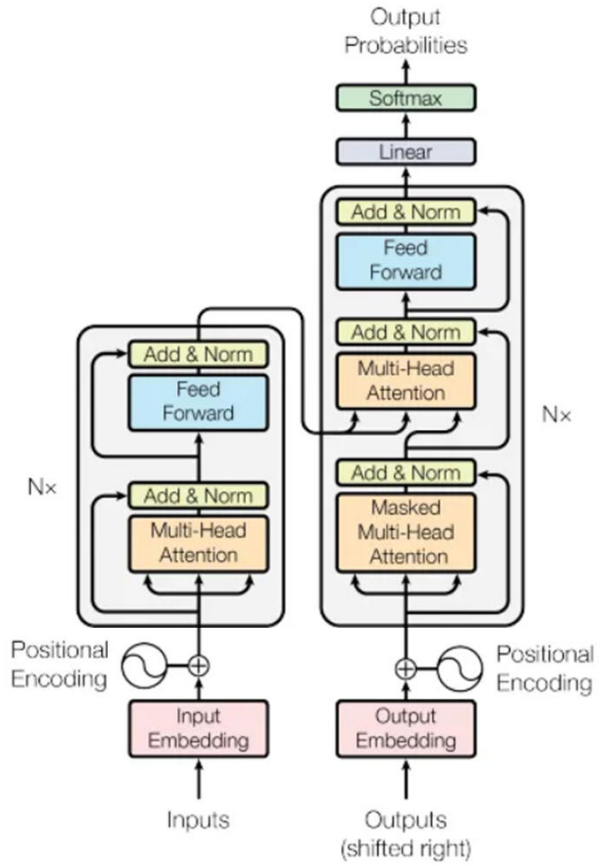
Attention Please

Attention Mechanism in Generative Models

Daniel Cohen-Or

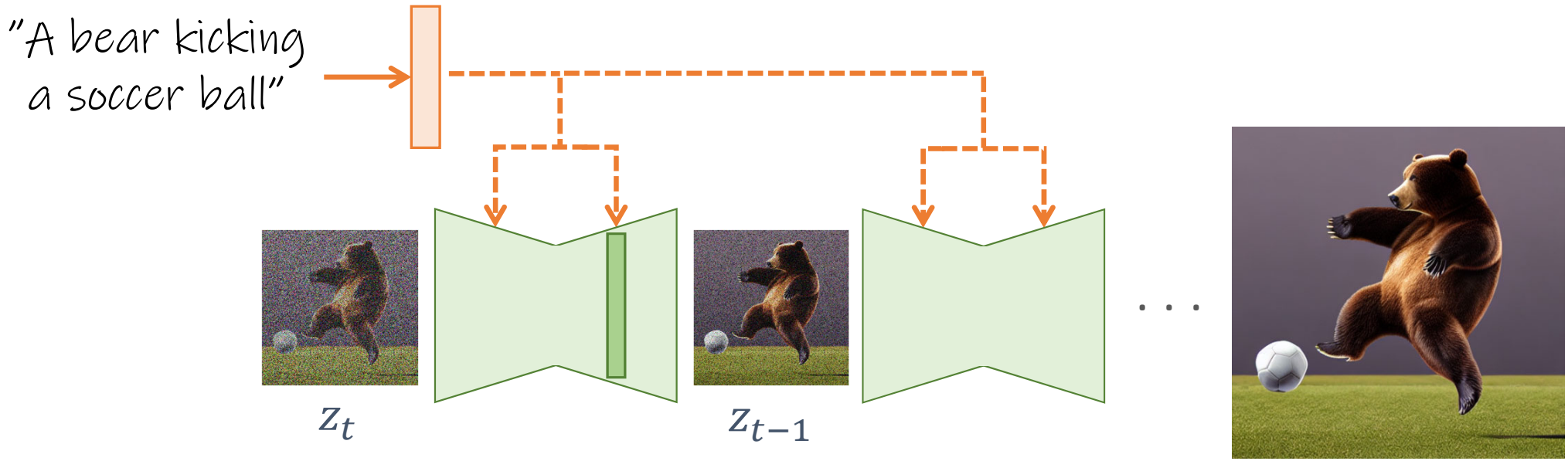
“Attention is all you need” , Waswani et al. 2007

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



From “Attention is all you need” paper by Vaswani, et al., 2017 [1]

The Denoising Process with Attention Layers



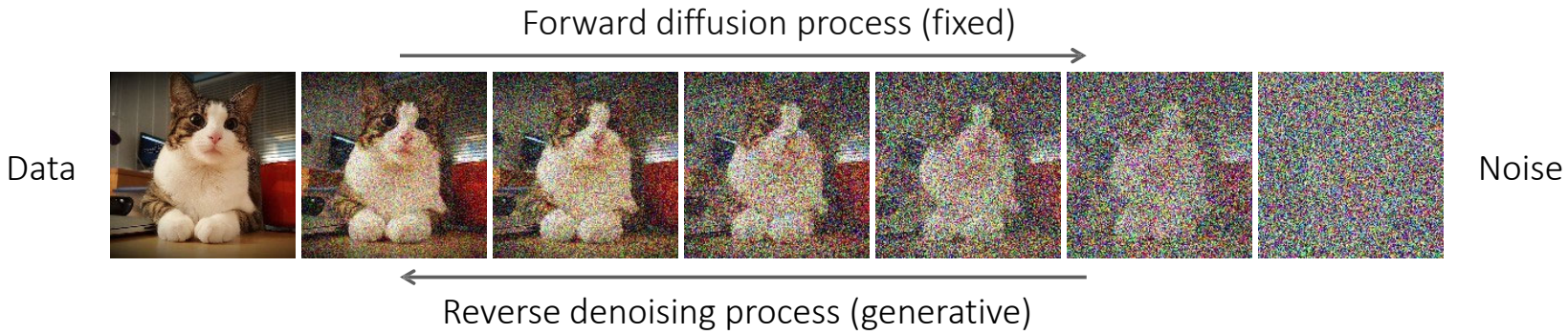
Denoising network

Denoising Diffusion Models

Learning to generate by denoising

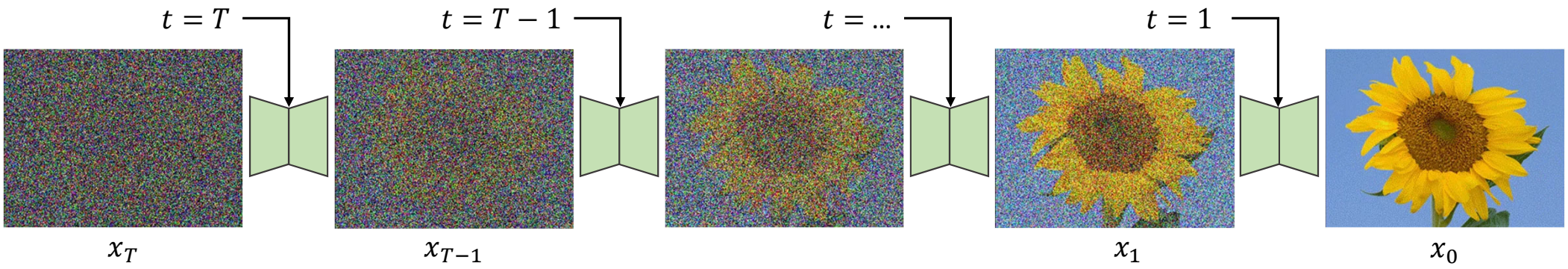
Denoising diffusion models consist of two processes:

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising

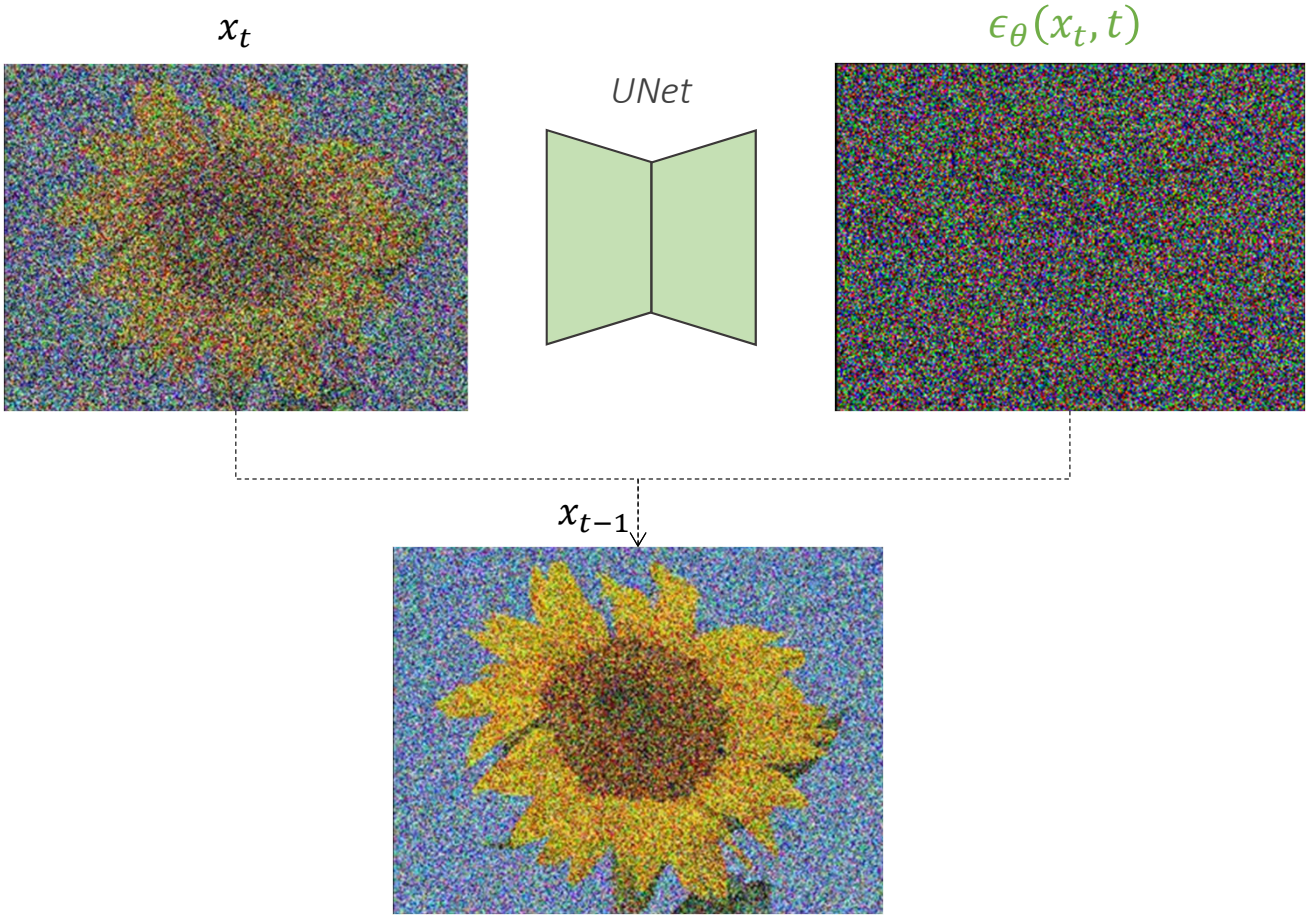


Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015 Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020
Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

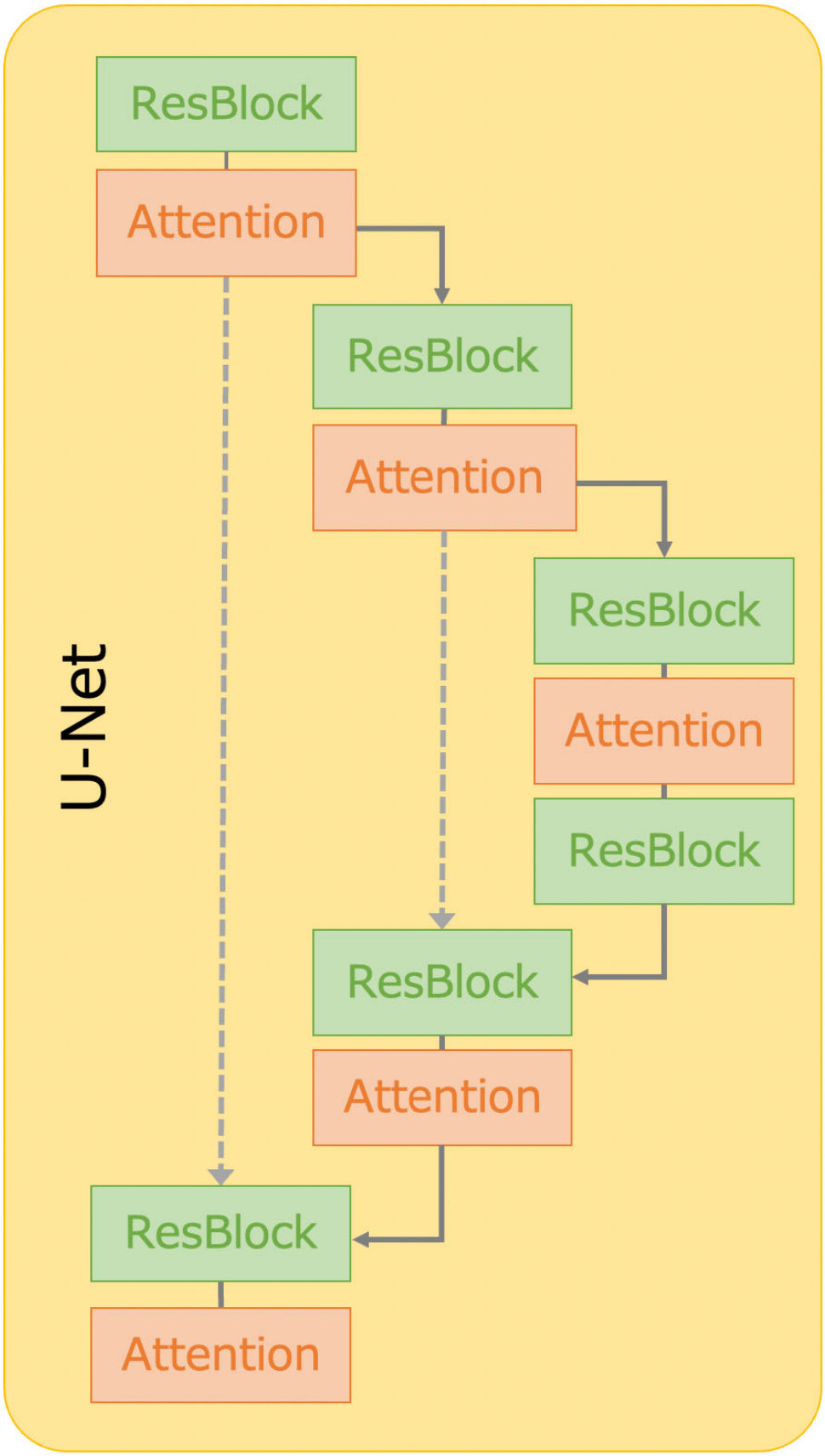
Diffusion Models

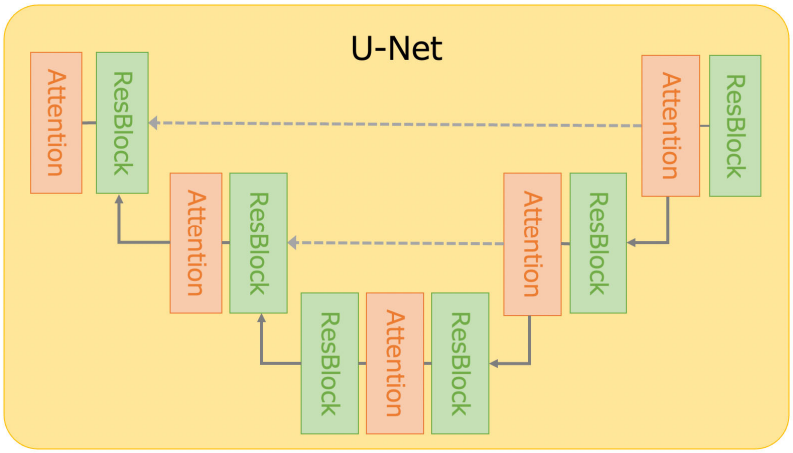


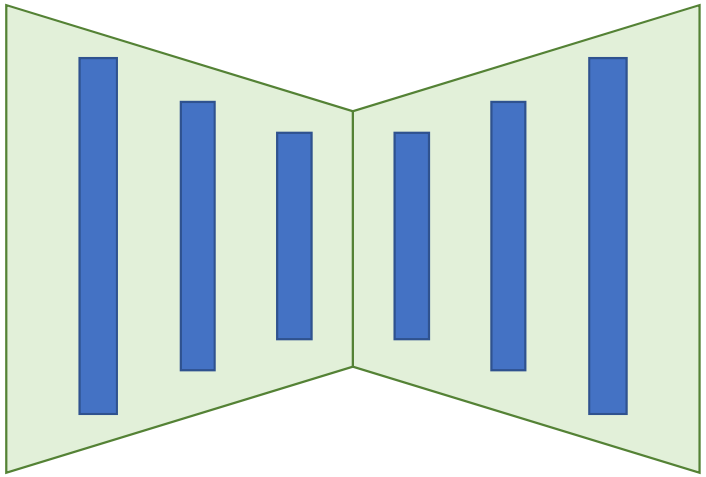
Diffusion Models



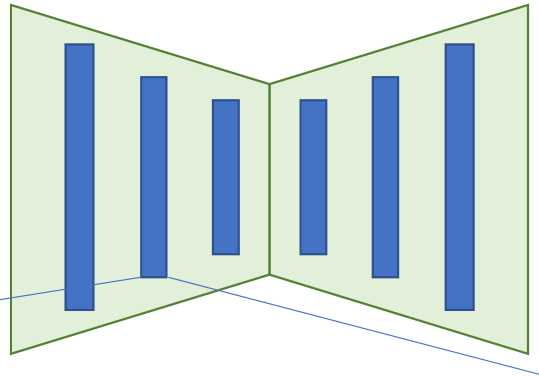
Unet

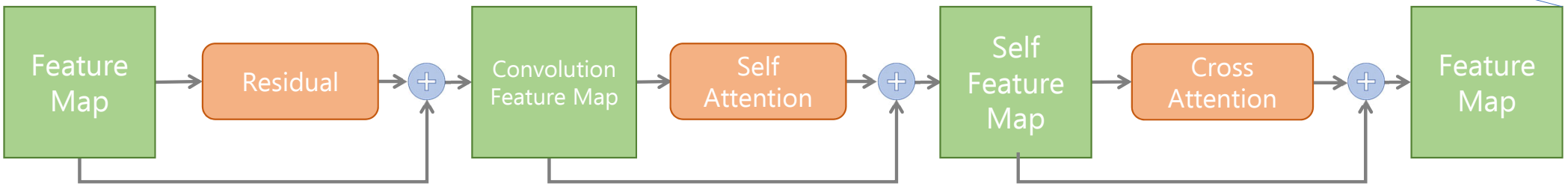

U-Net
ResBlock
Attention
ResBlock
Attention
ResBlock
Attention
ResBlock
ResBlock
Attention
ResBlock
Attention


U-Net
Attention
ResBlock
Attention
ResBlock
Attention
ResBlock
Attention
ResBlock
ResBlock
Attention
ResBlock
Attention
ResBlock



The UNet Layer

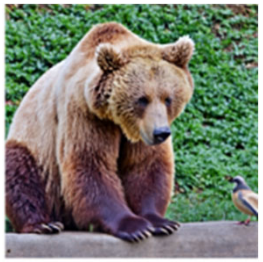




Cross-Attention

$N = 6$

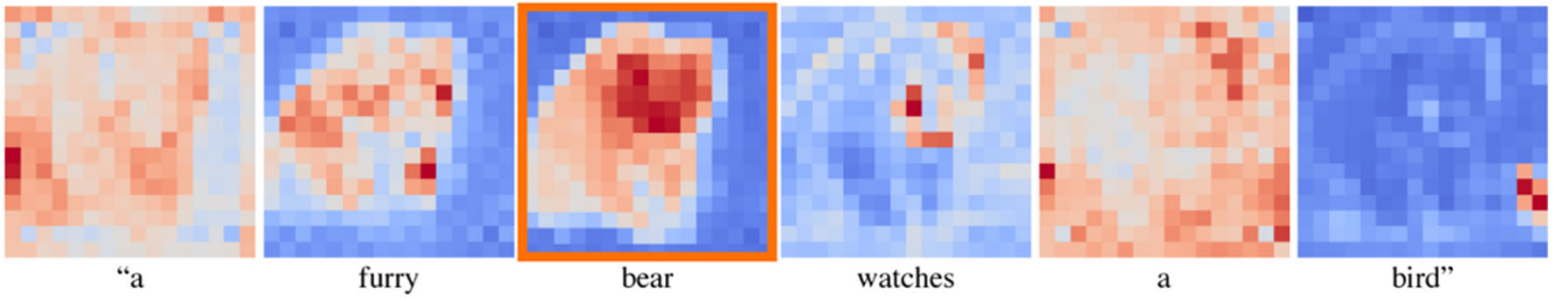
P^2



A
furry
bear
watches
a
bird

Q K

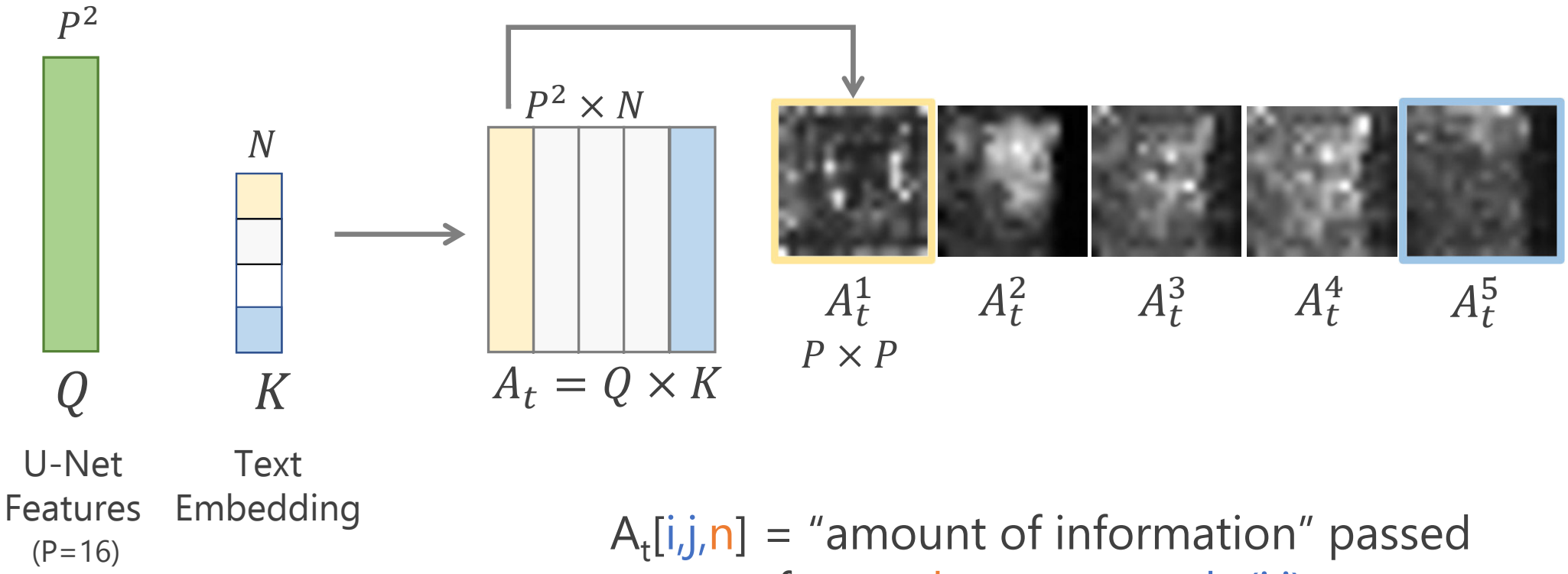
$P^2 \times N$



$$A_t = Q \times K$$

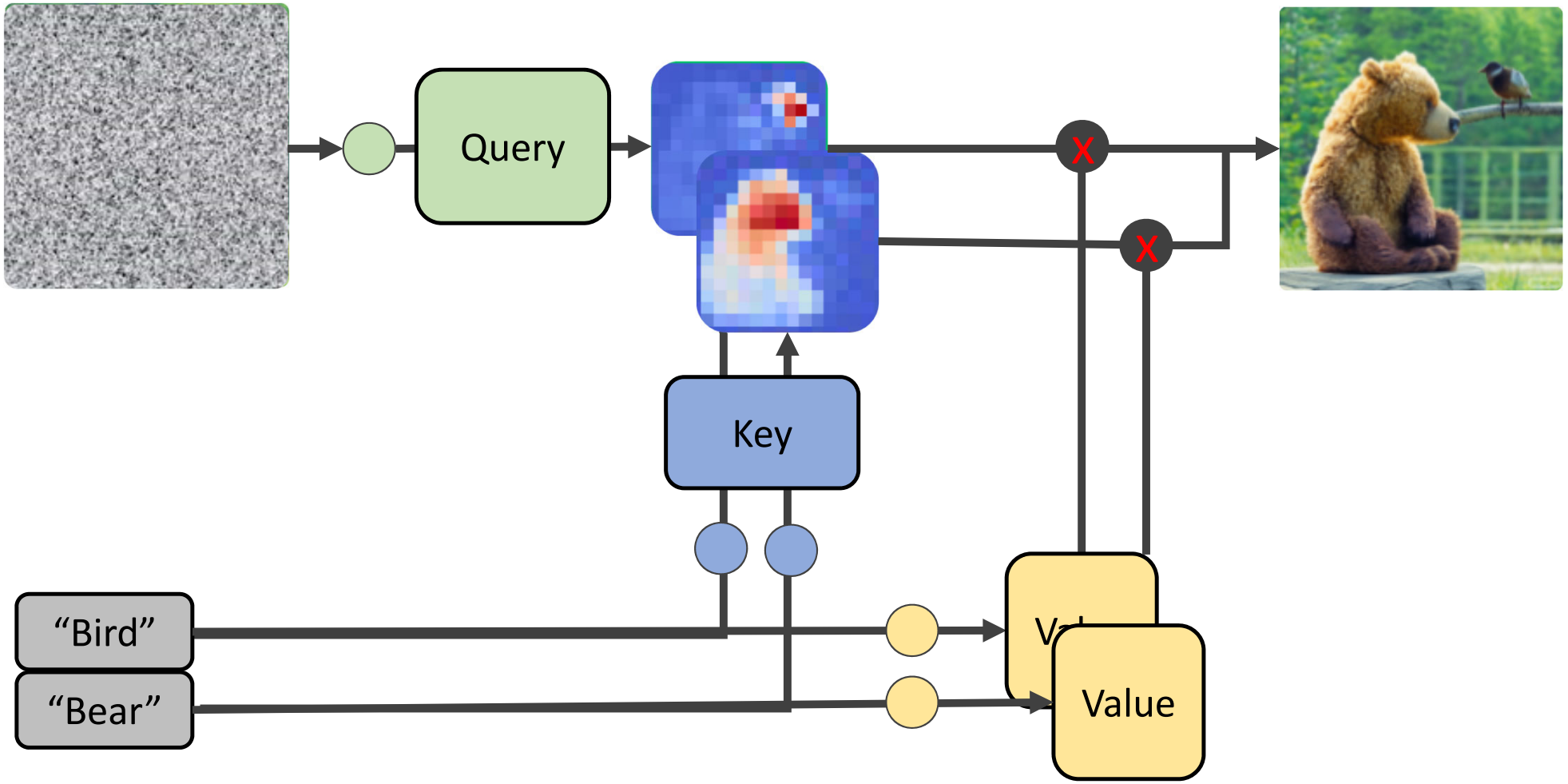
A_t[i,j,n] = "amount of information" passed from token n to patch (i,j)

The Cross-Attention

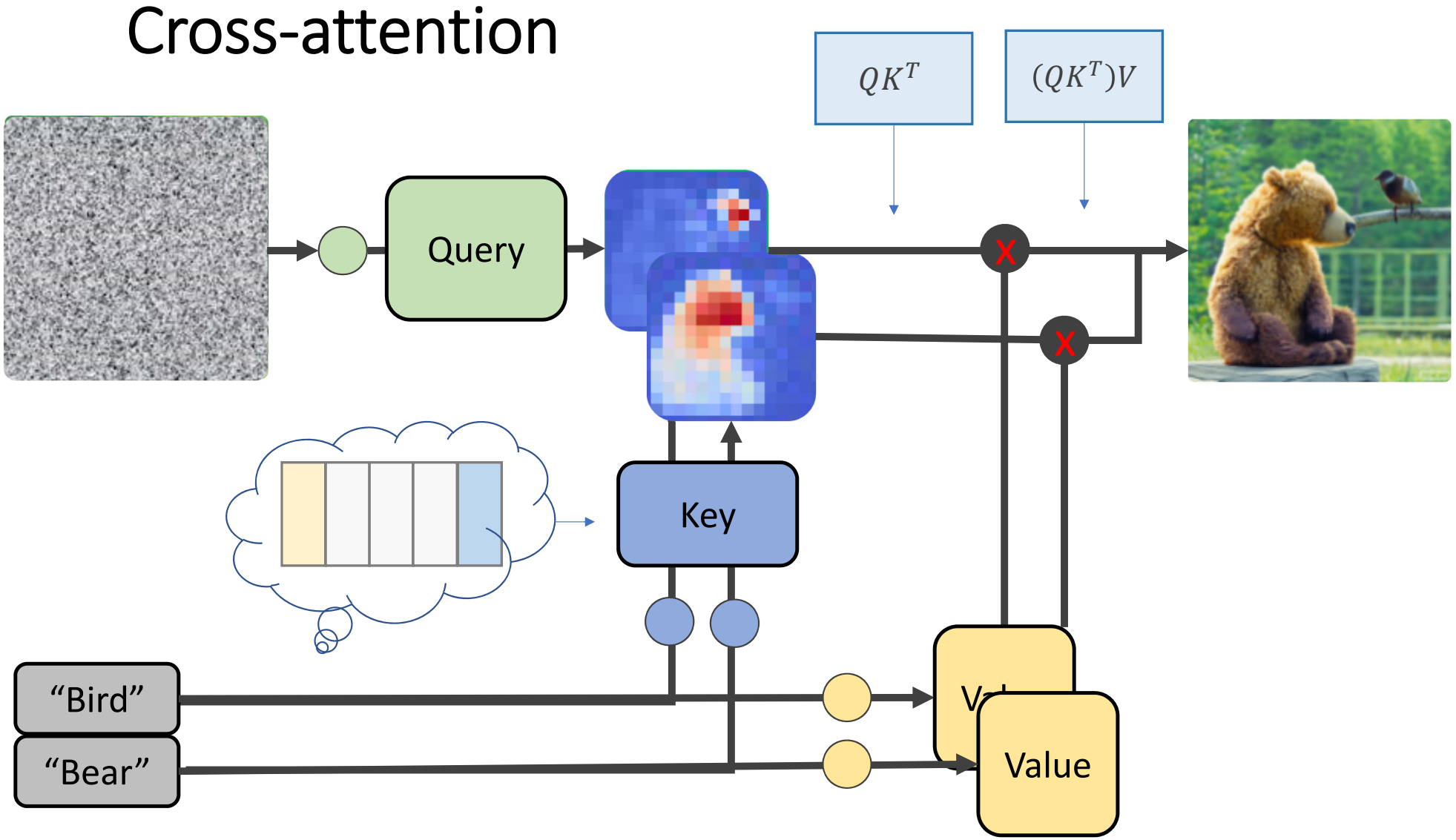


A_t[i,j,n] = "amount of information" passed from token n to patch (i,j)

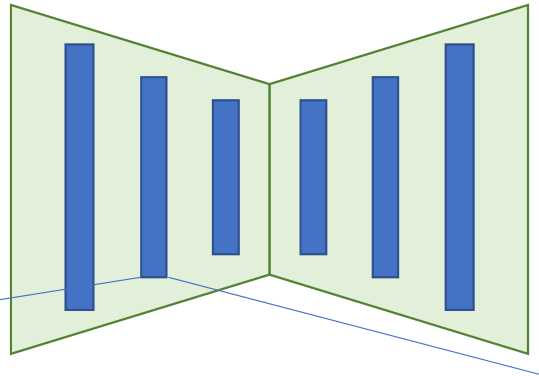
Cross-attention

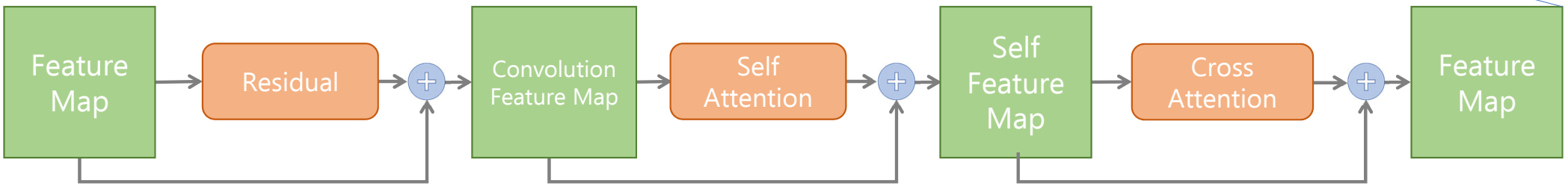


Cross-attention



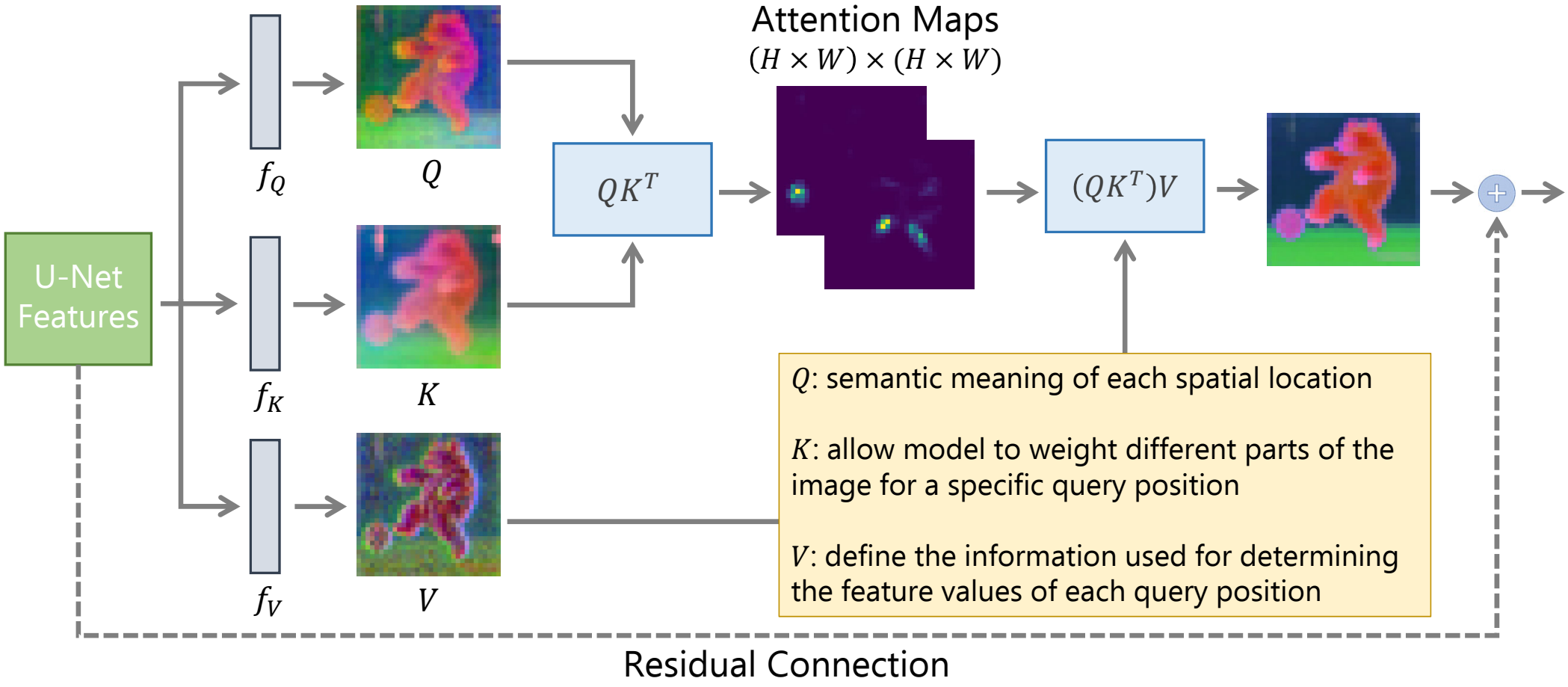
The UNet Layer



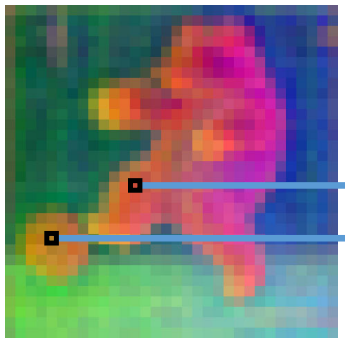


The Self-Attention

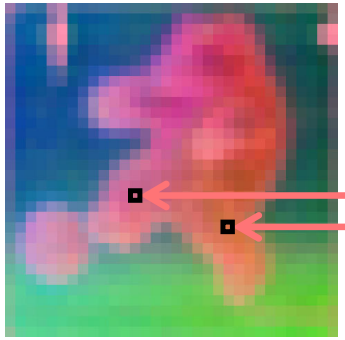
Represents where the model "looks" in the image for each spatial position in Q



The Self-Attention

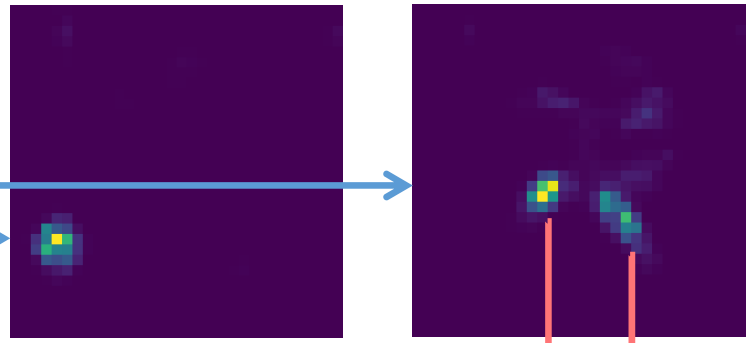


Queries
$H \times W \times C$



Keys
$H \times W \times C$

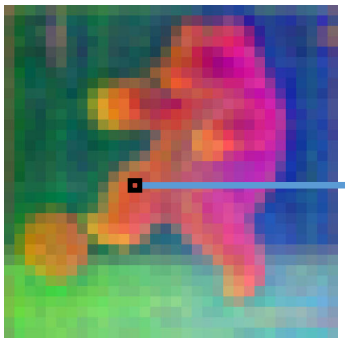
Attention Maps
$(H \times W) \times (H \times W)$



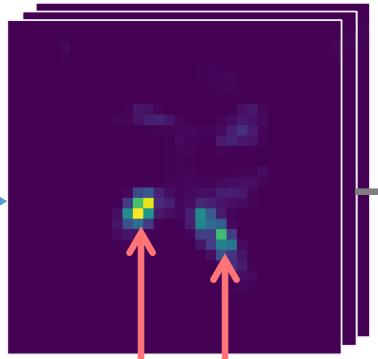
Each query defines to a
$H \times W$ attention map!

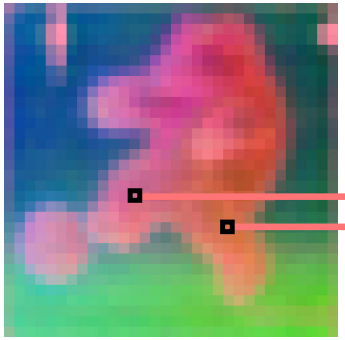
A **query** on the leg of the bear "attends"
to **keys** located on the leg of the bear!

The Self-Attention

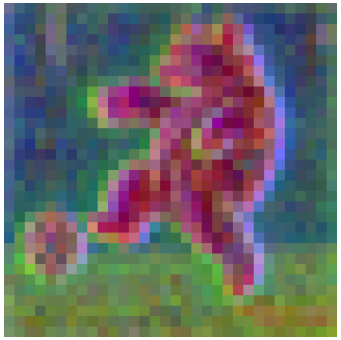


Queries
$H \times W \times C$





Keys
$H \times W \times C$



Values
$H \times W \times C$

Values
$H \times W \times C$

$(QK^T)V$

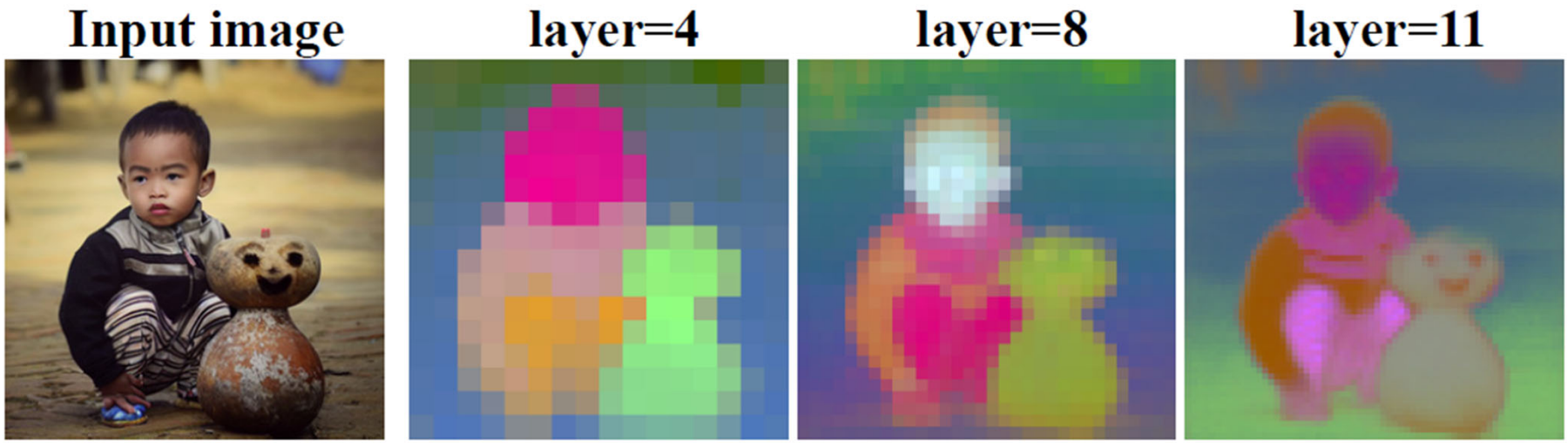
t = 0.6, layer: 10 / 70

t = 0.6, layer: 35 / 70

Self-Segmentation

Localizing Object-level Shape Variations [Patashnik et al., ICCV 2023]

Self-Attention Maps



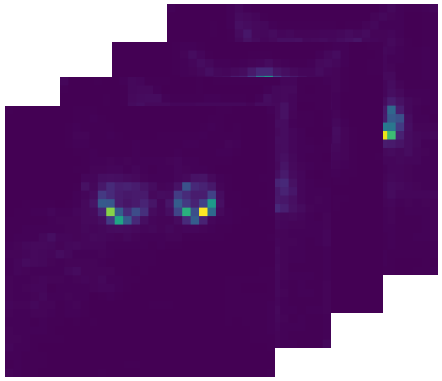
Are these PCA on the self-attention ? On what exactly the QK maps?

Plug-and-Play [Tumanyan et al., CVPR 2023]

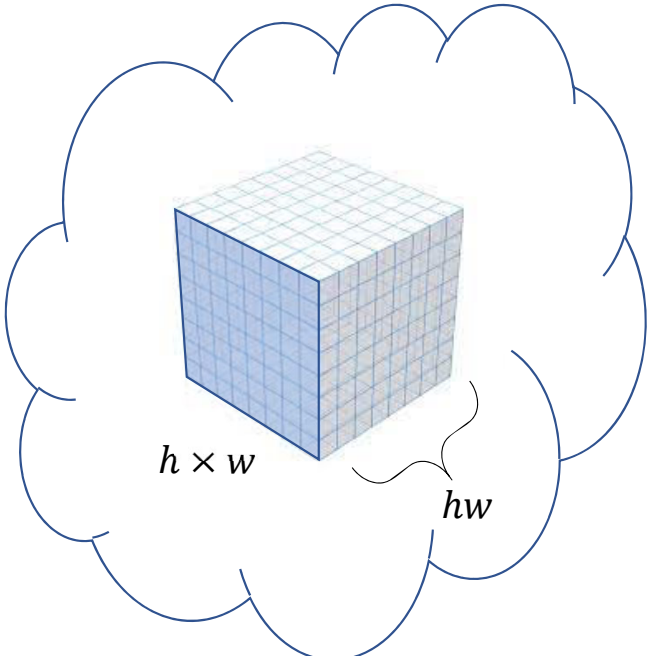
Self-Segmentation

"a **cat** is wearing sunglasses"

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



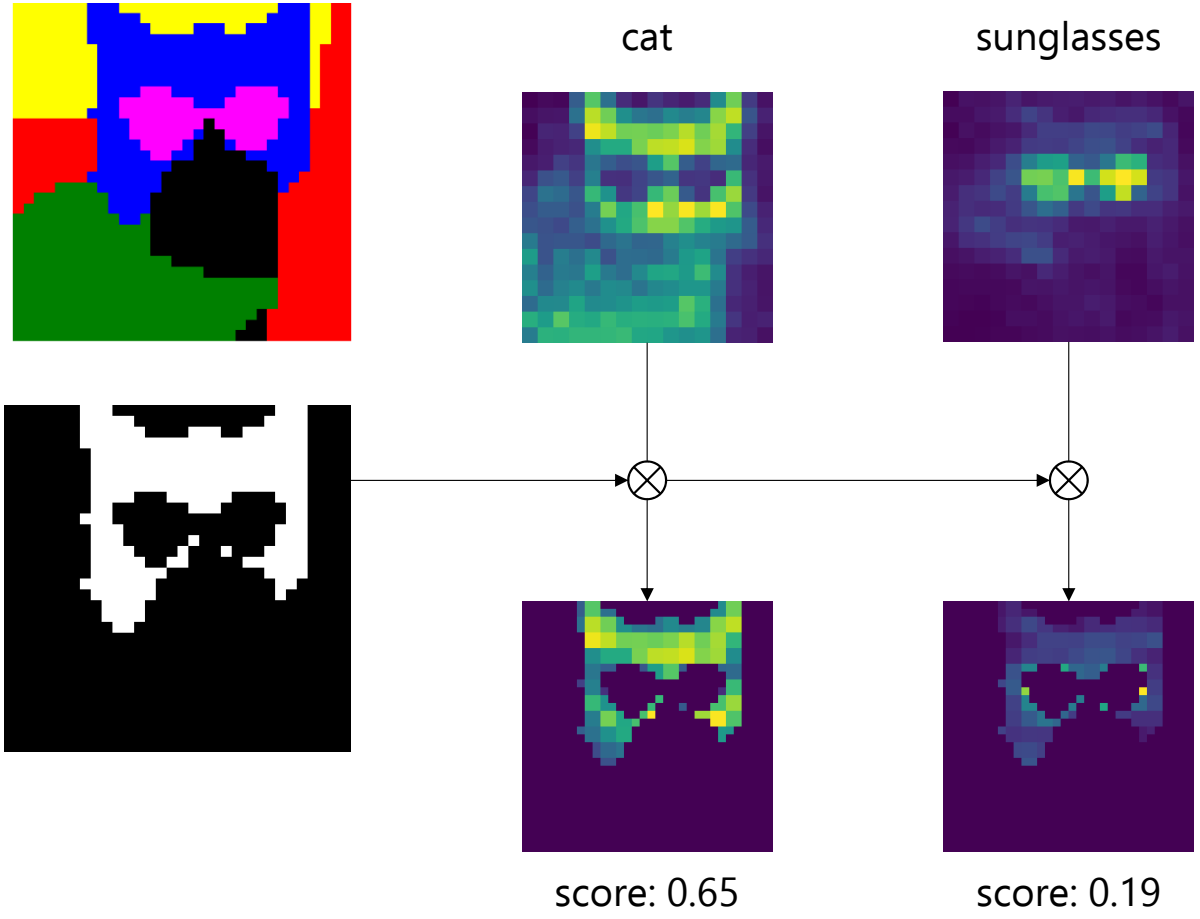
$hw \times (h \times w)$



cluster

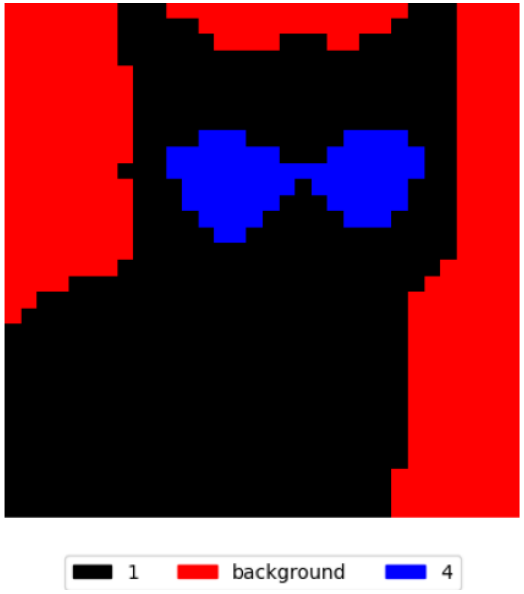
Segments labeling

“a cat is wearing sunglasses”



Segments labeling

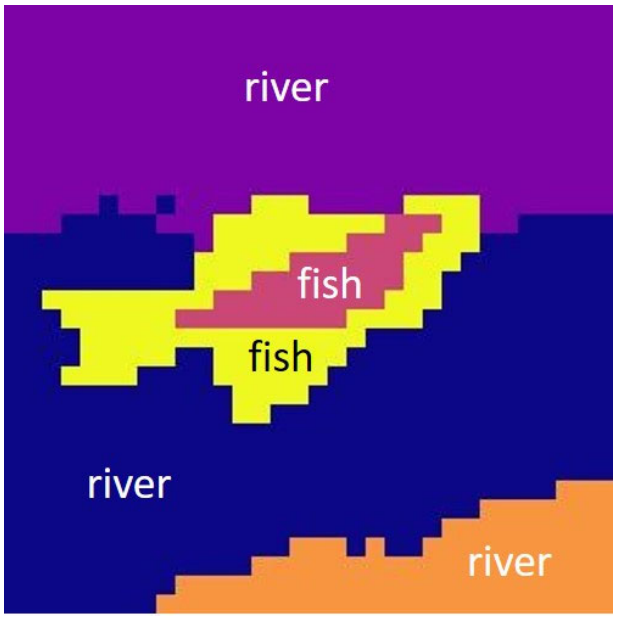
“a cat is wearing sunglasses”



1-cat, 4-sunglasses

Self-Segmentation Results

Self-Segmentation Results



Cross-Image Attention

Cross-Image Attention for Zero-Shot Appearance Transfer

Yuval Alaluf*, Daniel Garibi*, Or Patashnik, Hadar Averbuch Elor, Daniel Cohen-Or

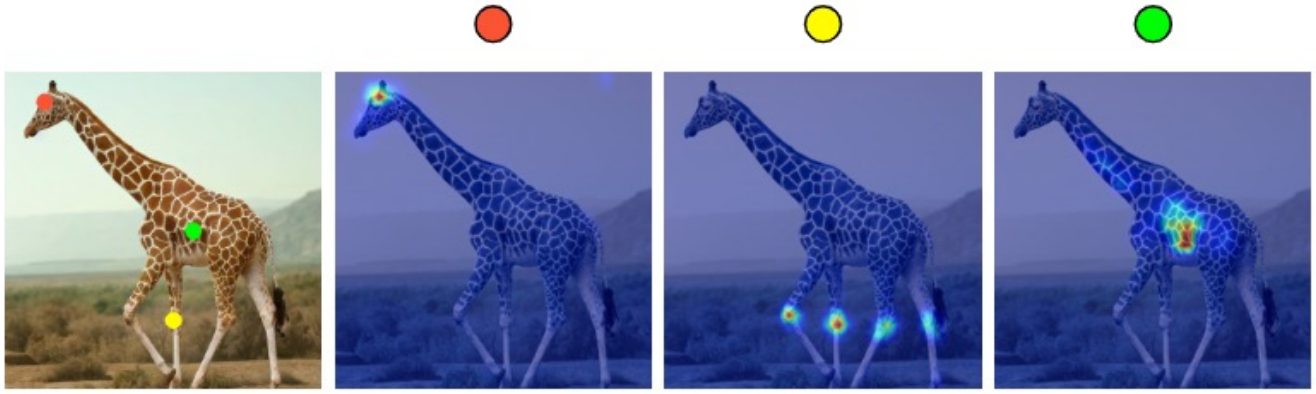
Motivation

Structure

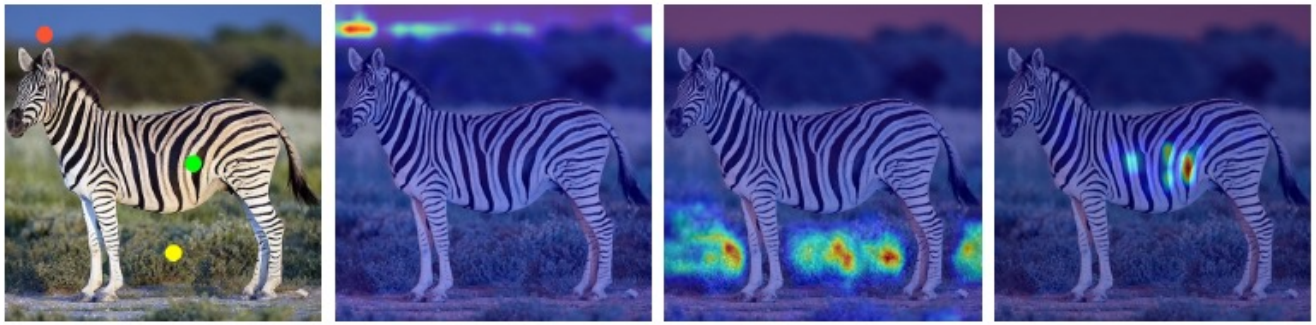
Appearance

Output

The Roles of the Queries, Keys, and Values



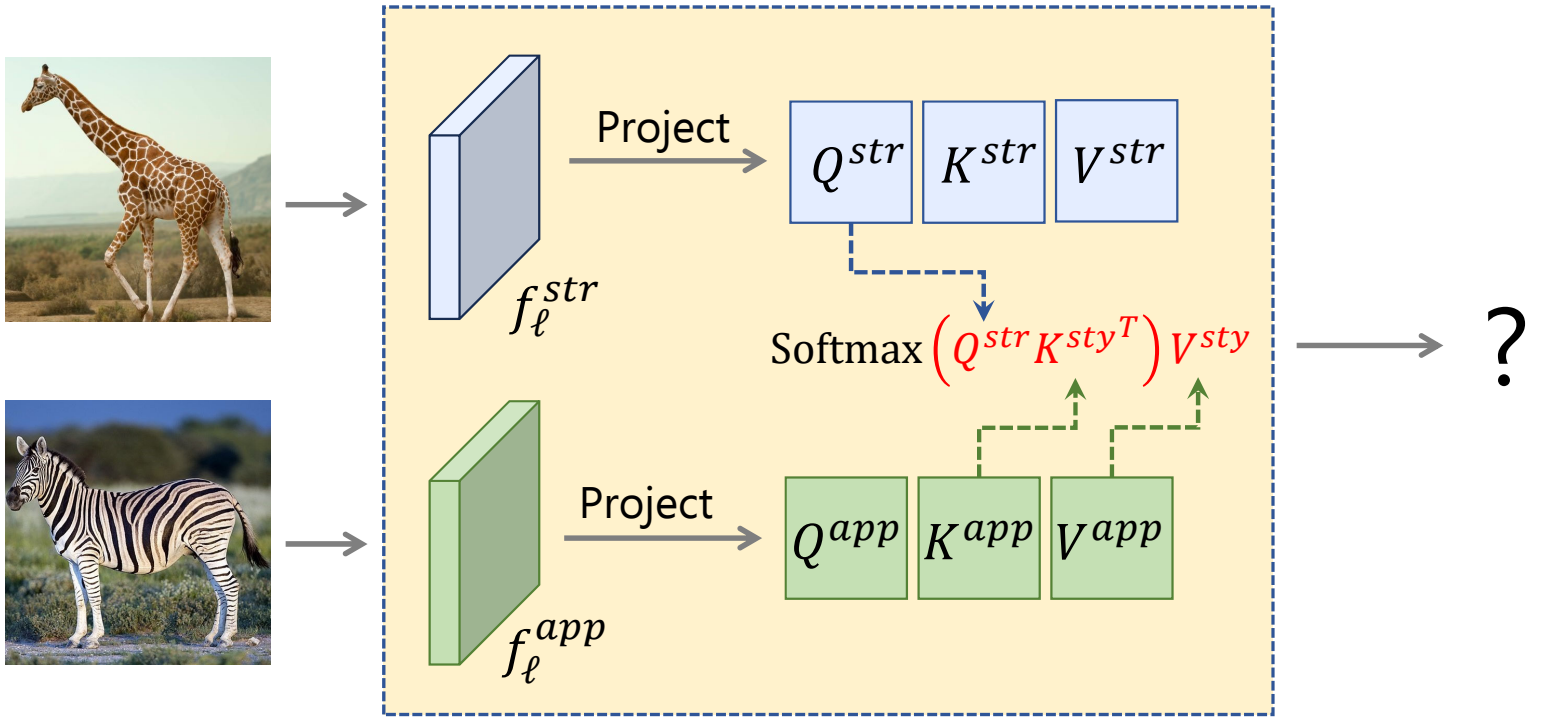
$$Q_{struct} \cdot K_{struct}^T$$



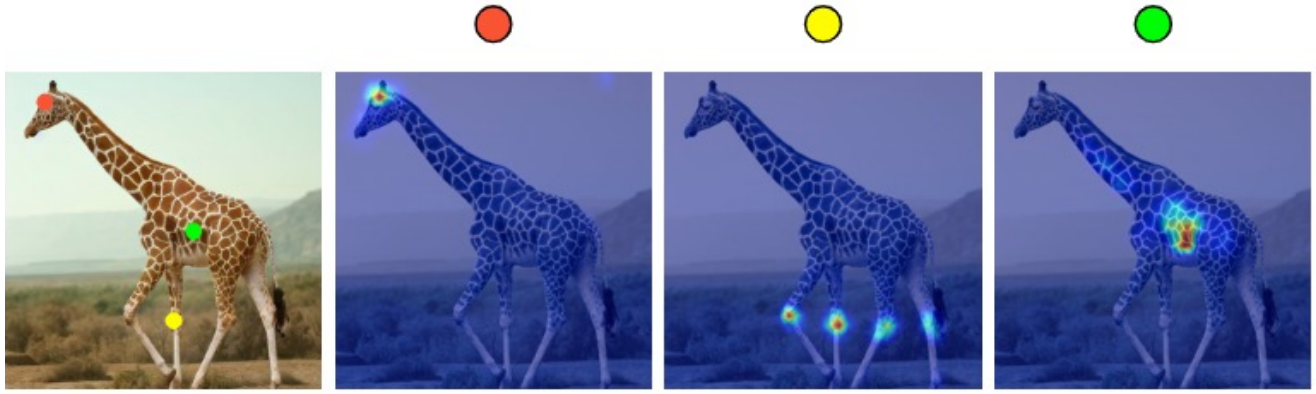
$$Q_{app} \cdot K_{app}^T$$

Self-attention maps, which focus on semantically similar regions in the image.

What If we Swapped the Queries, Keys, and Values Between Different Images?



The Roles of the Queries, Keys, and Values

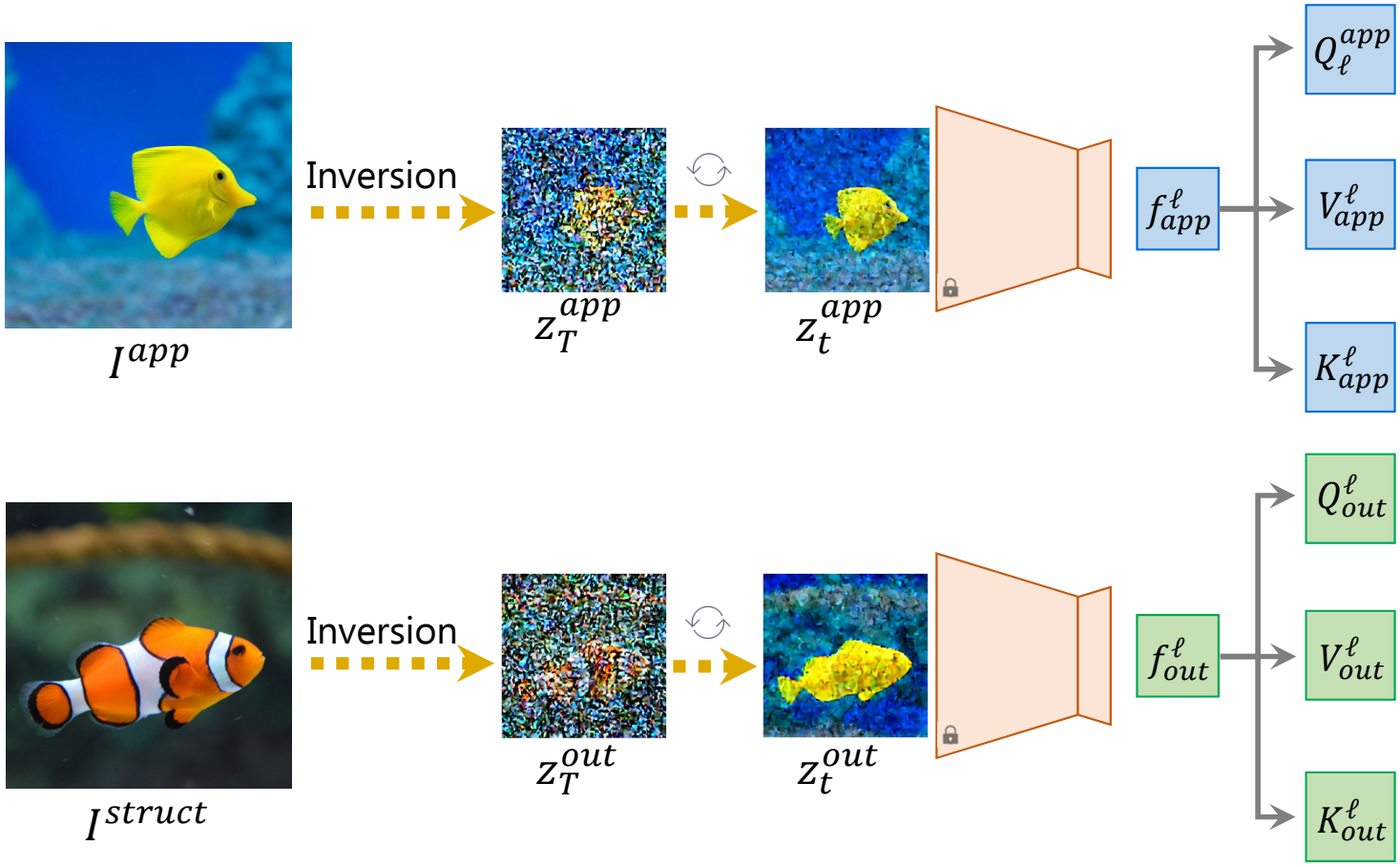


$$Q_{struct} \cdot K_{struct}^T$$

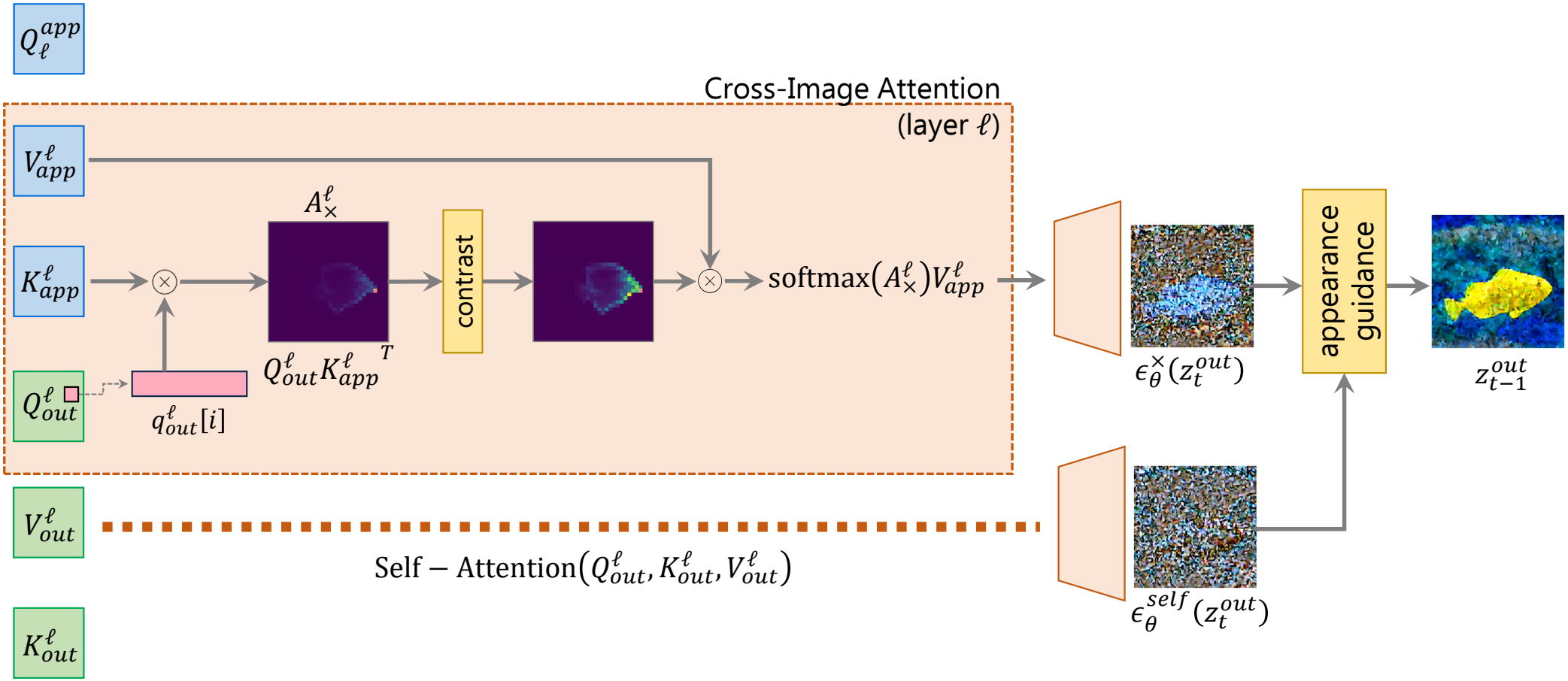
$$Q_{struct} \cdot K_{app}^T$$

Taking the **queries** from the structure image and the **keys** from the appearance image gives semantic correspondences between objects!

The Cross-Image Attention



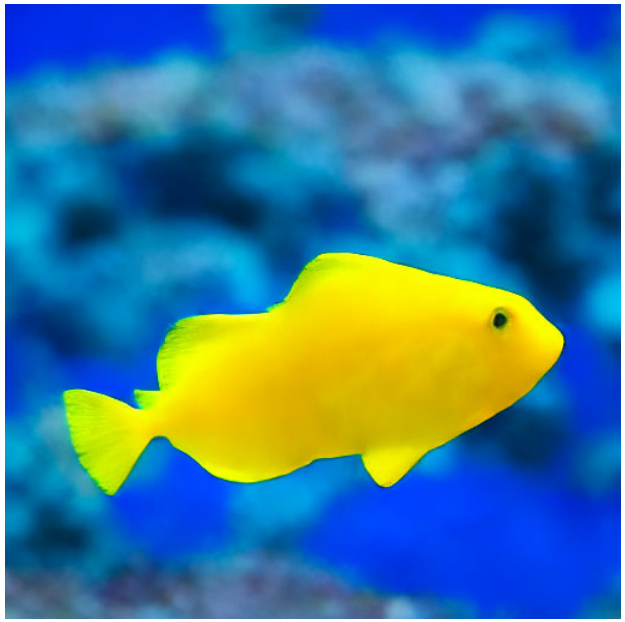
The Cross-Image Attention



The Cross-Image Attention

Structure

Appearance



Output

Appearance Transfer Results

Structure

Appearance

Output

Appearance Transfer Results

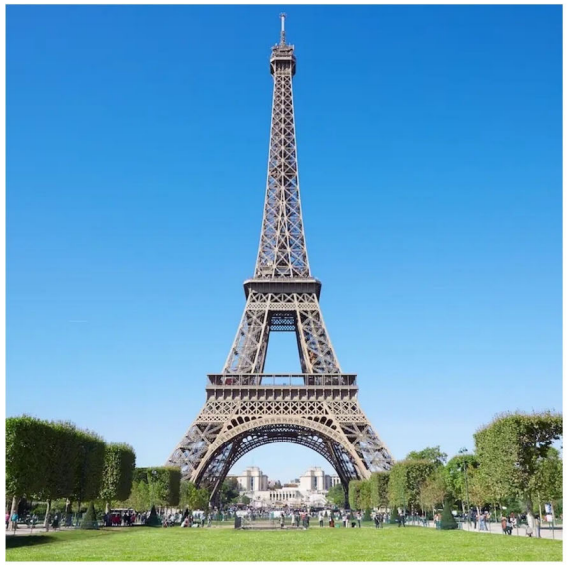
Structure

Appearance

Output

Appearance Transfer Results

Eiffel Tower



Structure

Sagrada Família

Appearance

Output

Appearance Transfer Results

Structure

Appearance

Output

Appearance Transfer Results

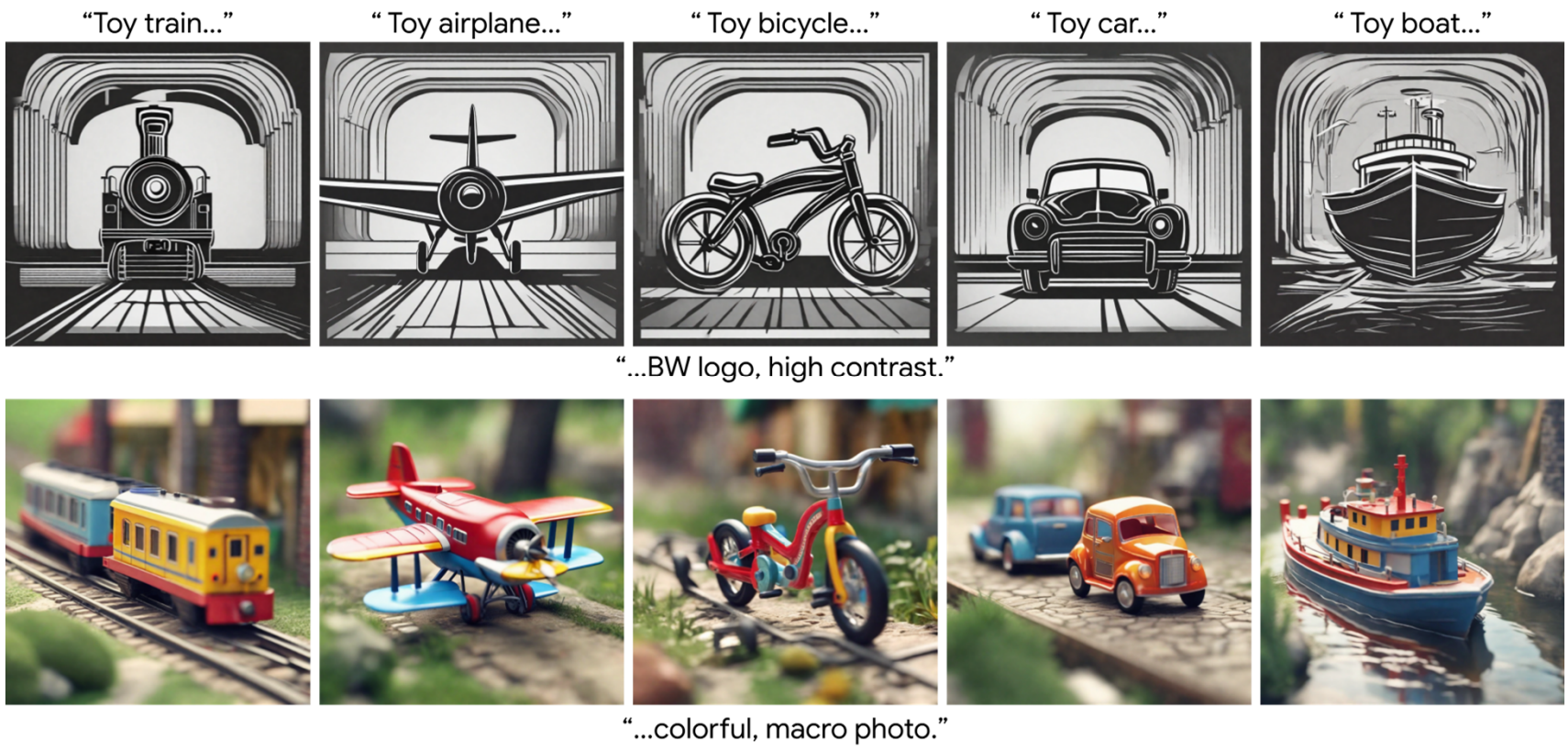
Structure

Appearance

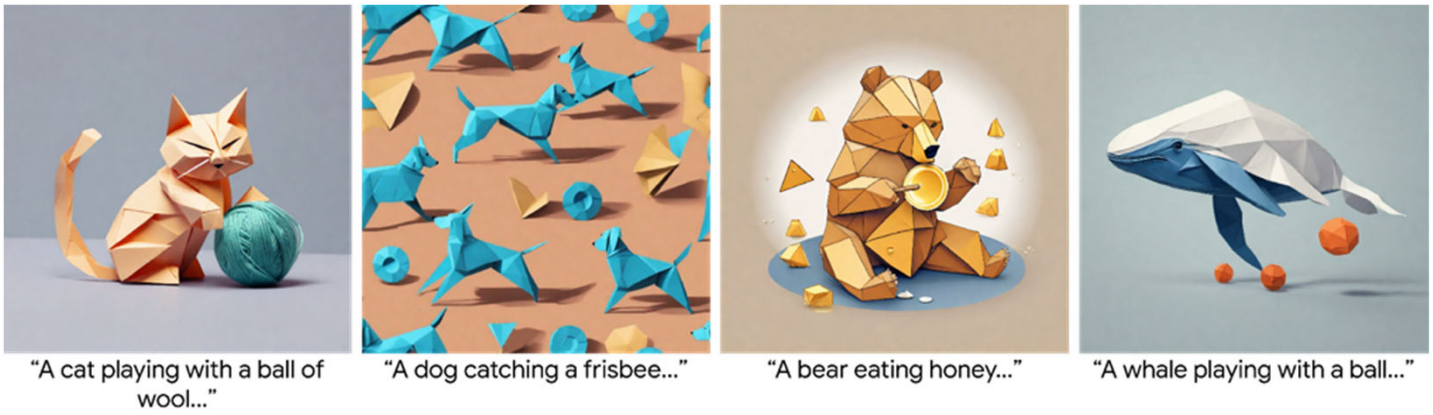
Output

StyleAlign

(link)



Text-to-Image Generation

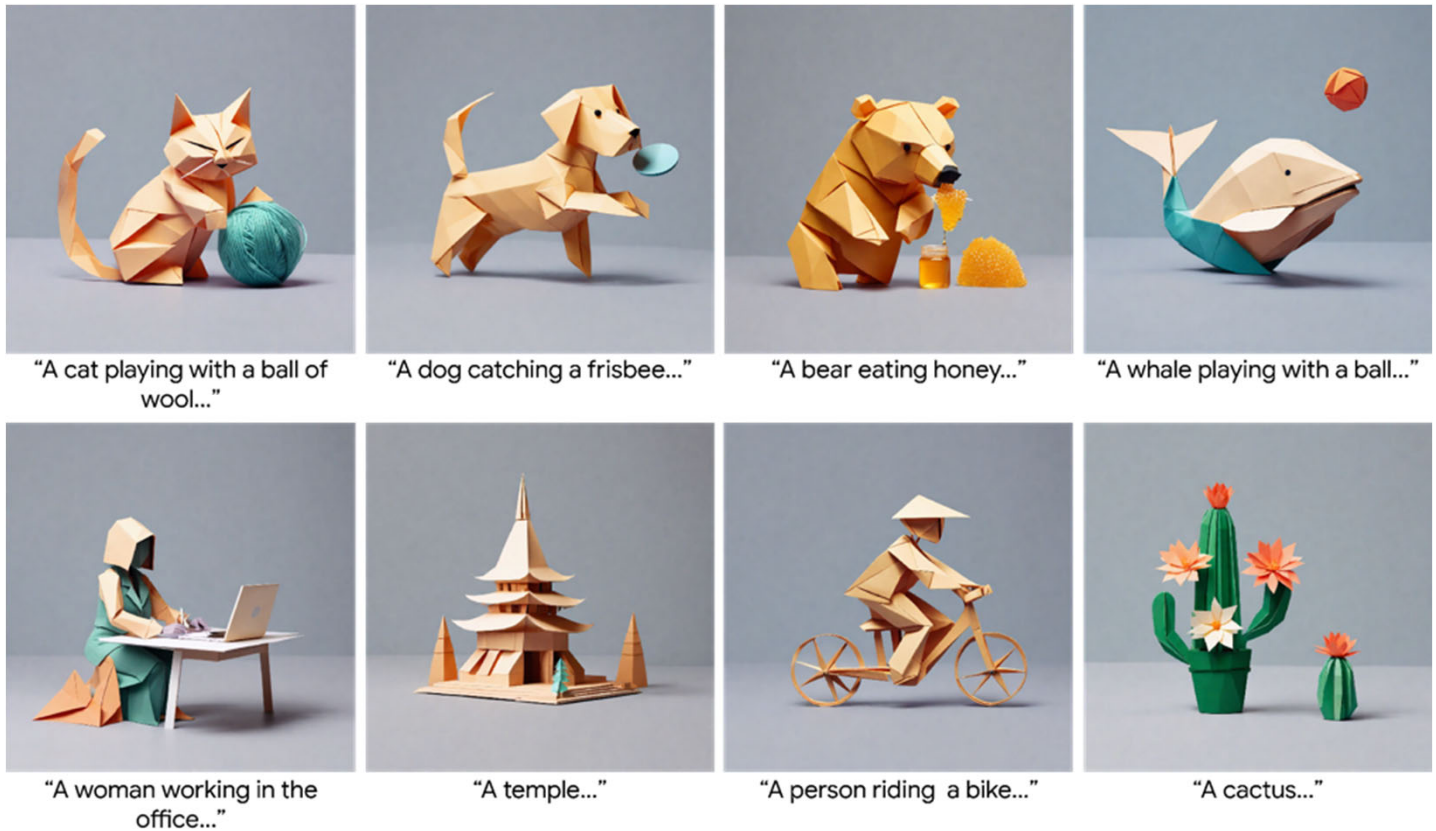


"A cat playing with a ball of wool..." "A dog catching a frisbee..." "A bear eating honey..." "A whale playing with a ball..."

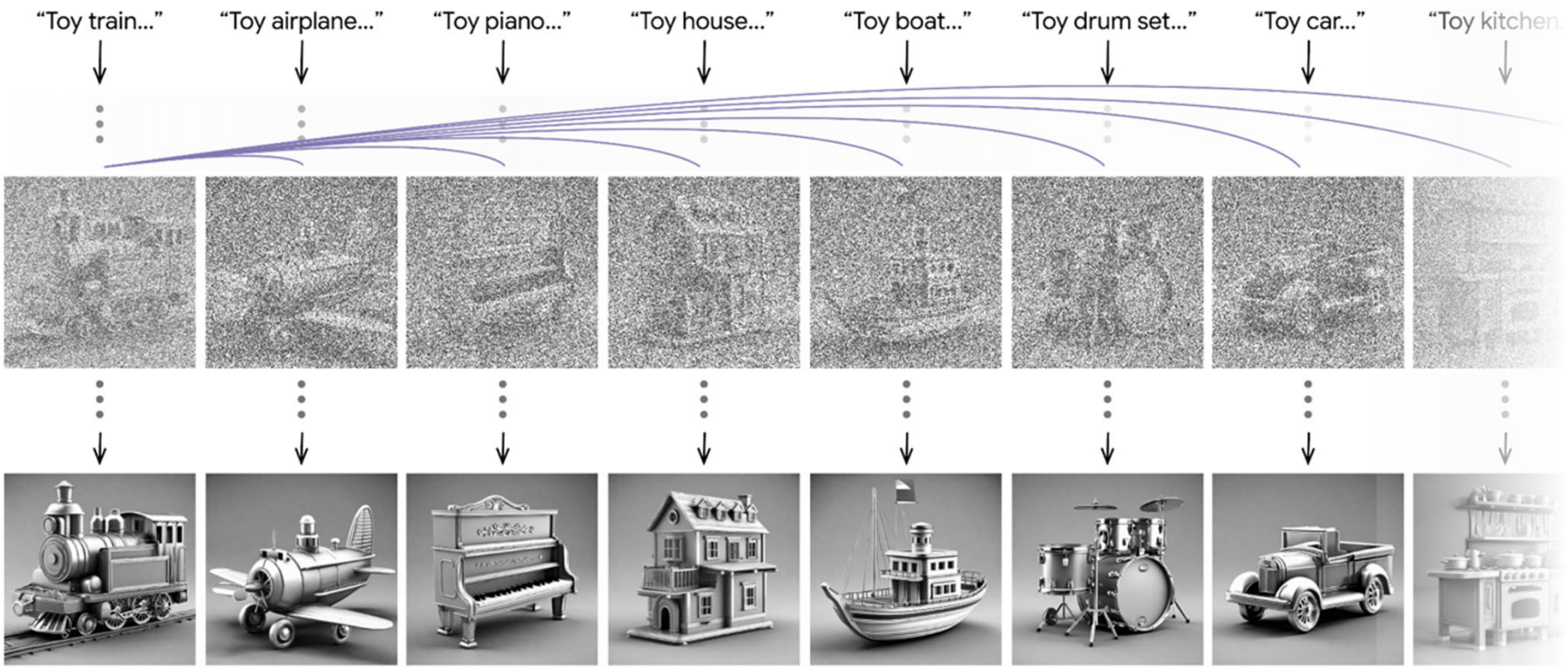
"A woman working in the office..." "A temple..." "A person riding a bike..." "A cactus..."

"... in minimal origami style."

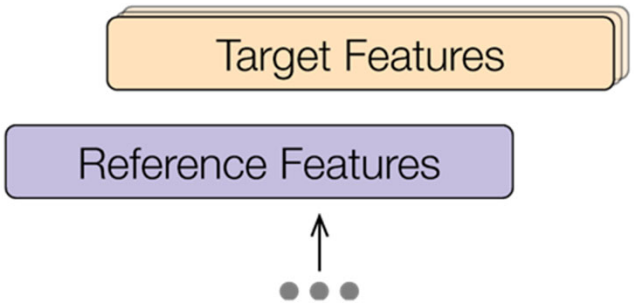
Text-to-Image Generation with StyleAligned



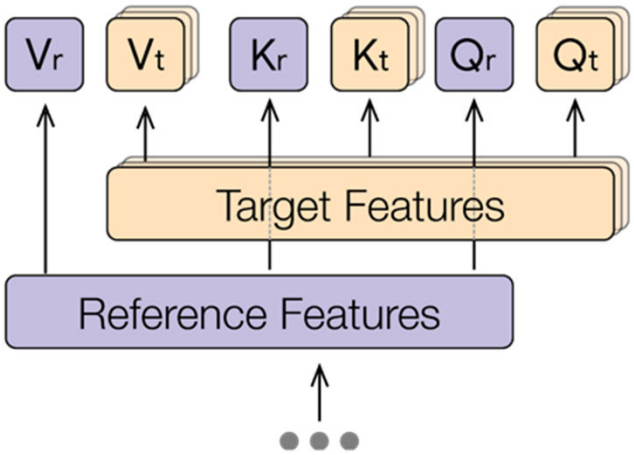
Shared attention during the diffusion process



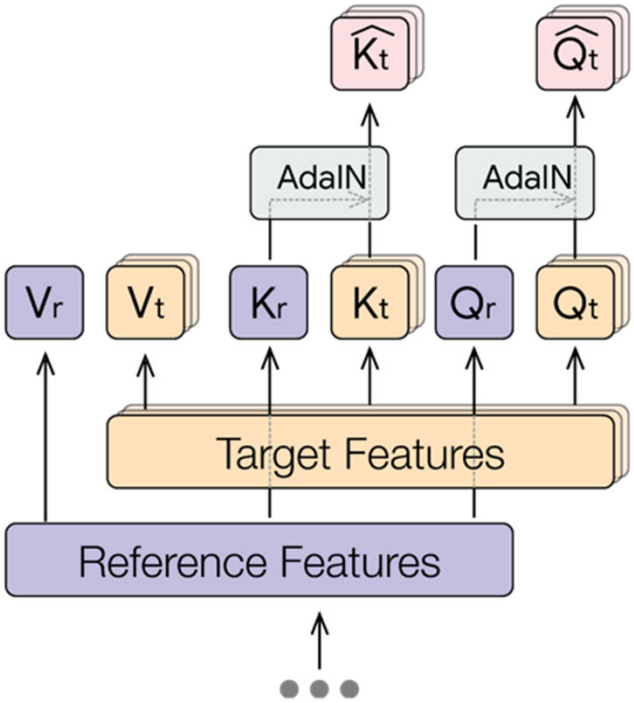
Shared Attention Layer



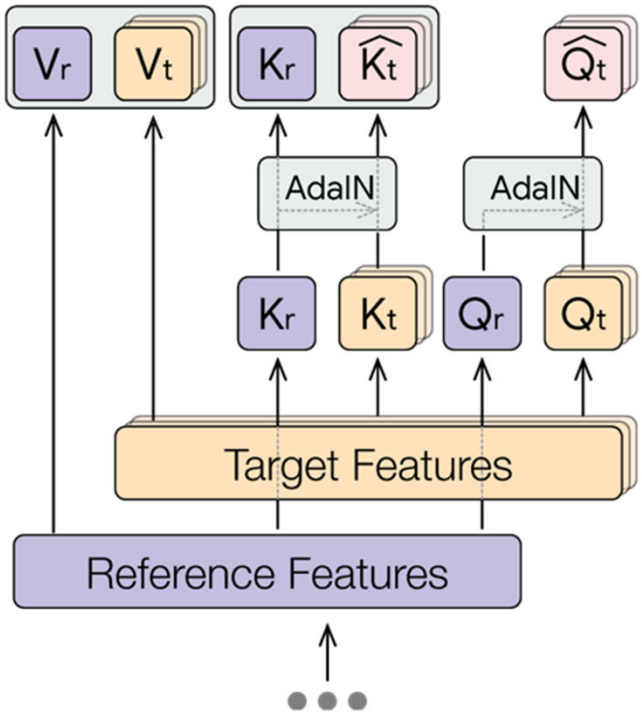
Shared Attention Layer



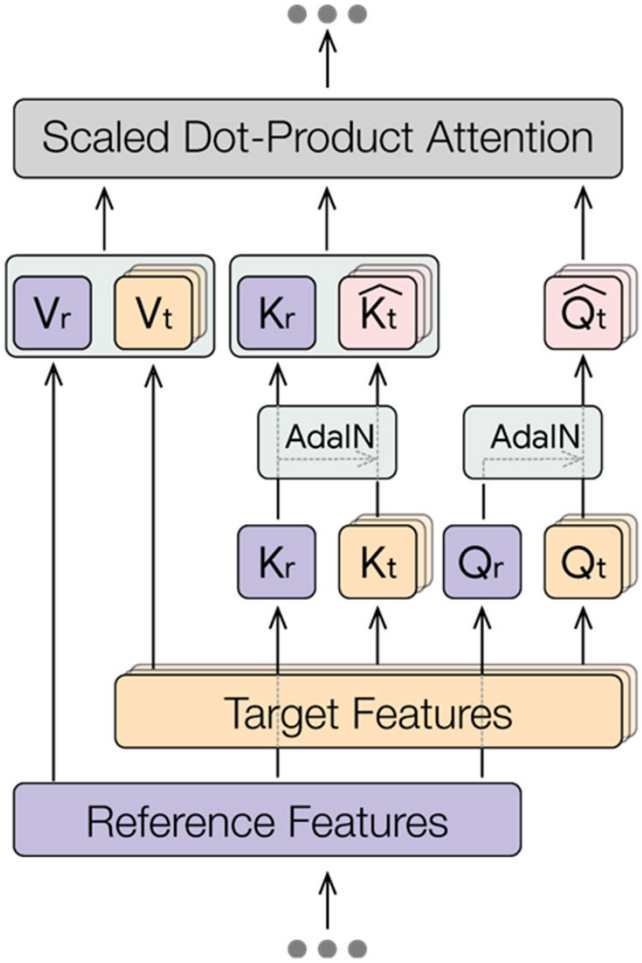
Shared Attention Layer



Shared Attention Layer

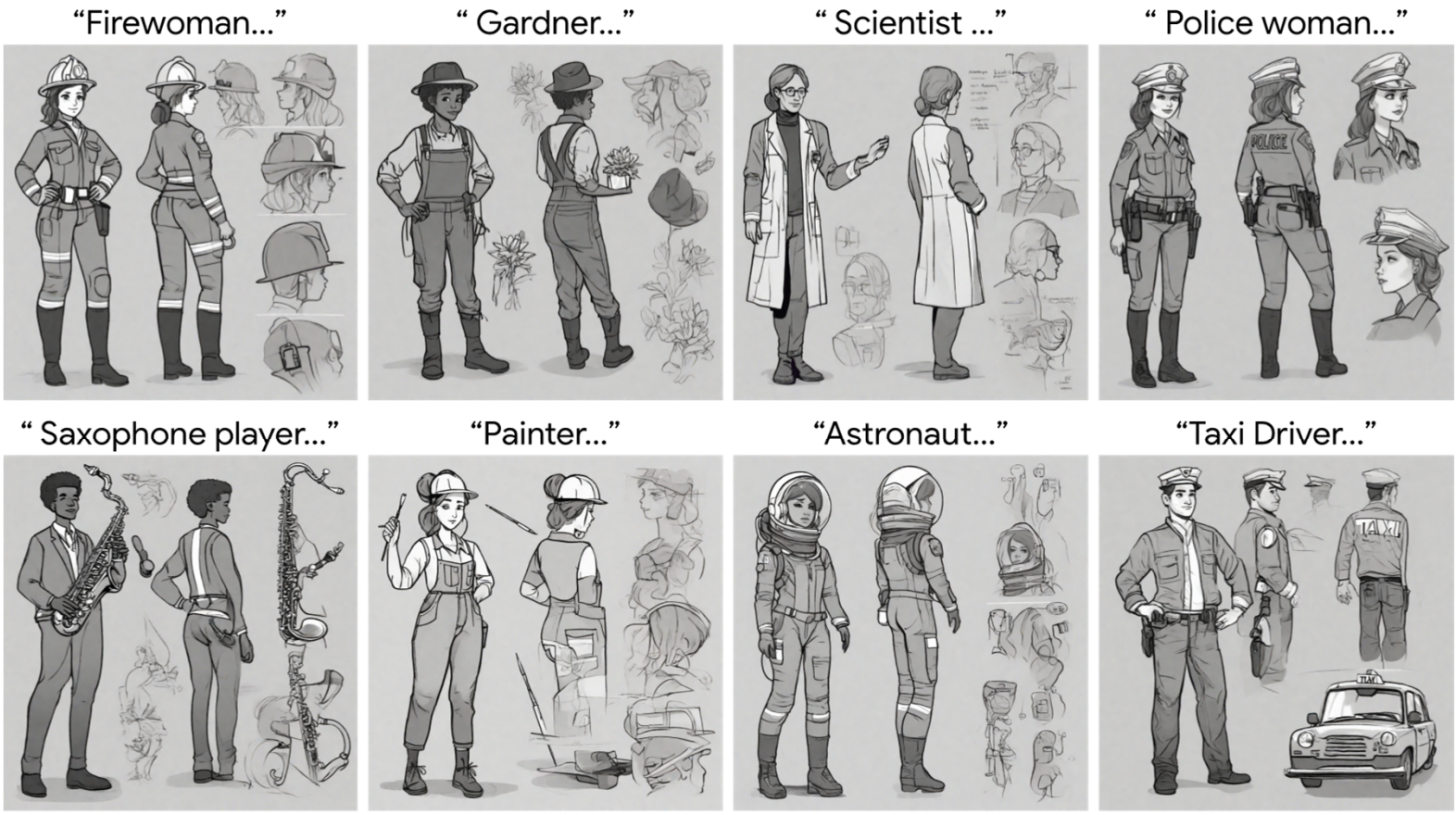


Shared Attention Layer



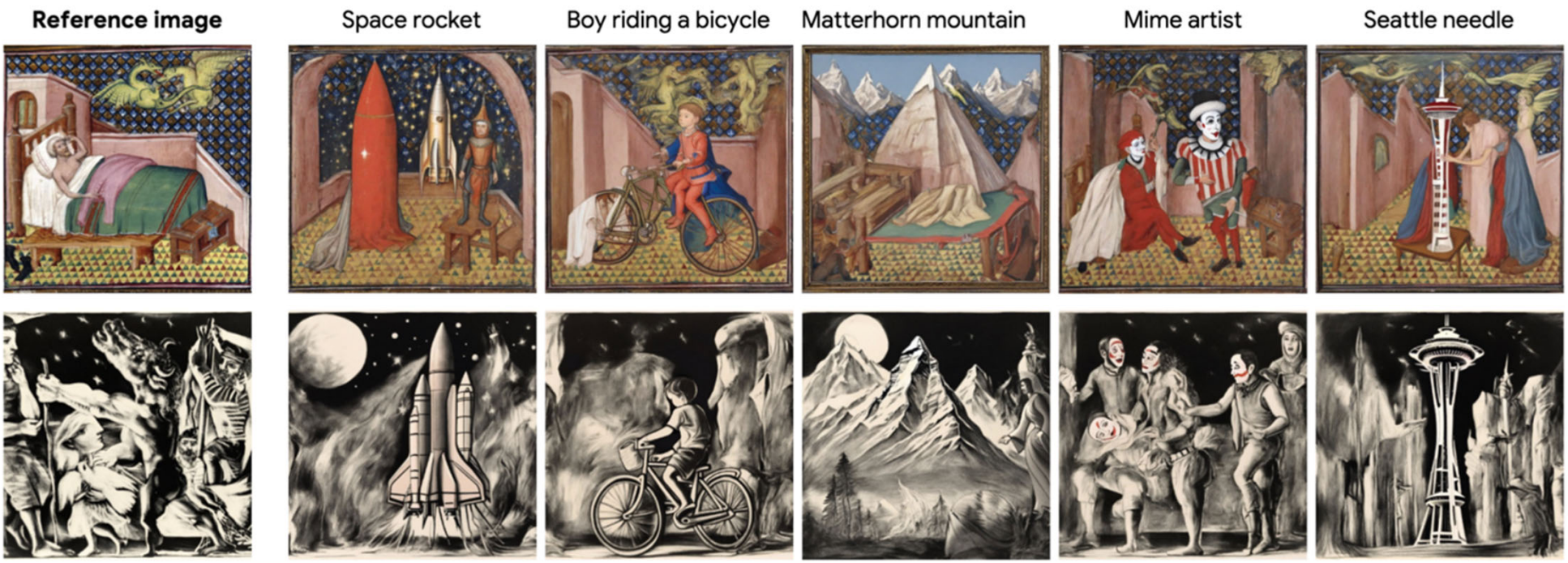
Style Aligned generation of Synthetic Images

Style Aligned generation of Synthetic Images



Style Aligned generation of Synthetic Images

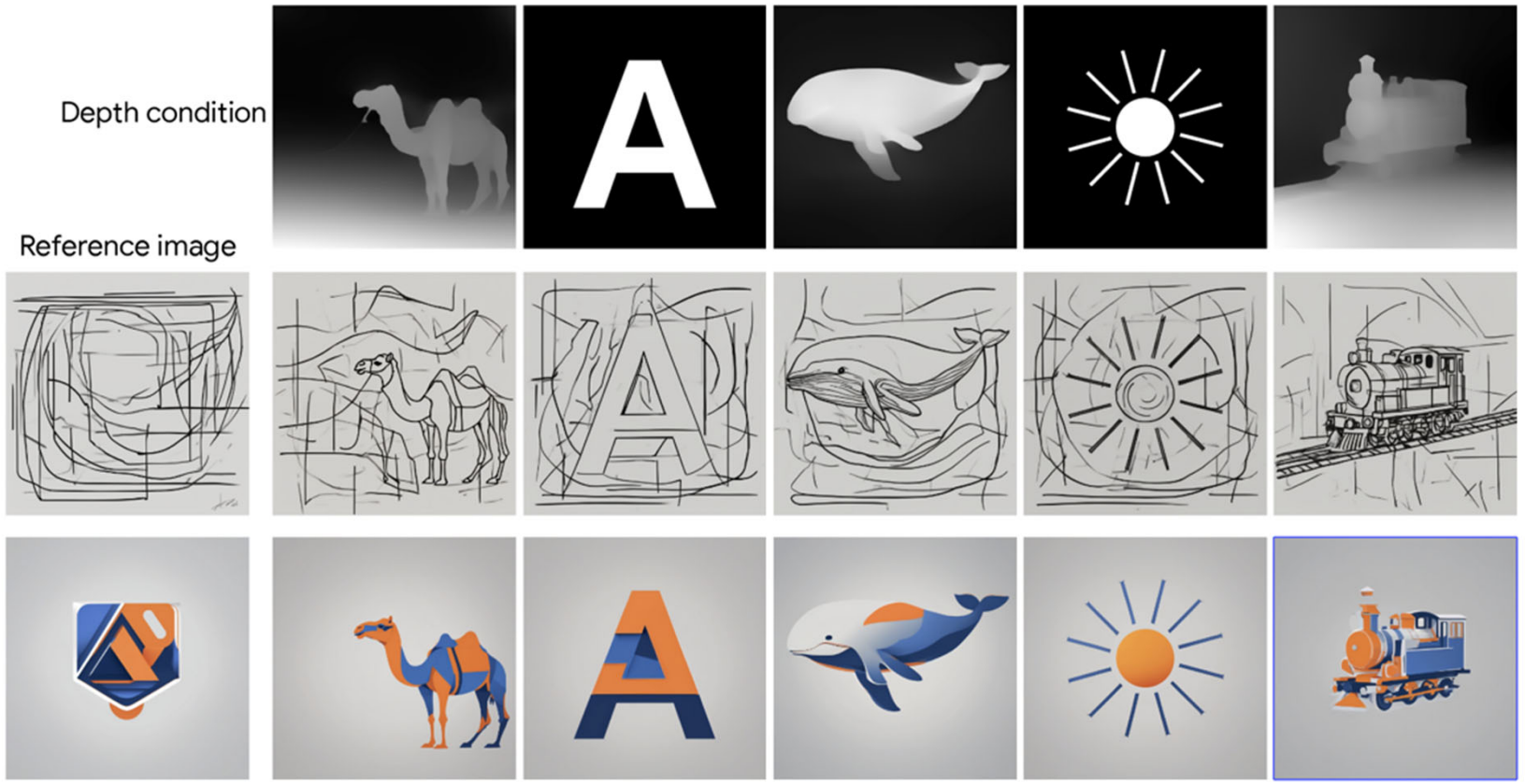
Style Aligned generation from an Input Image



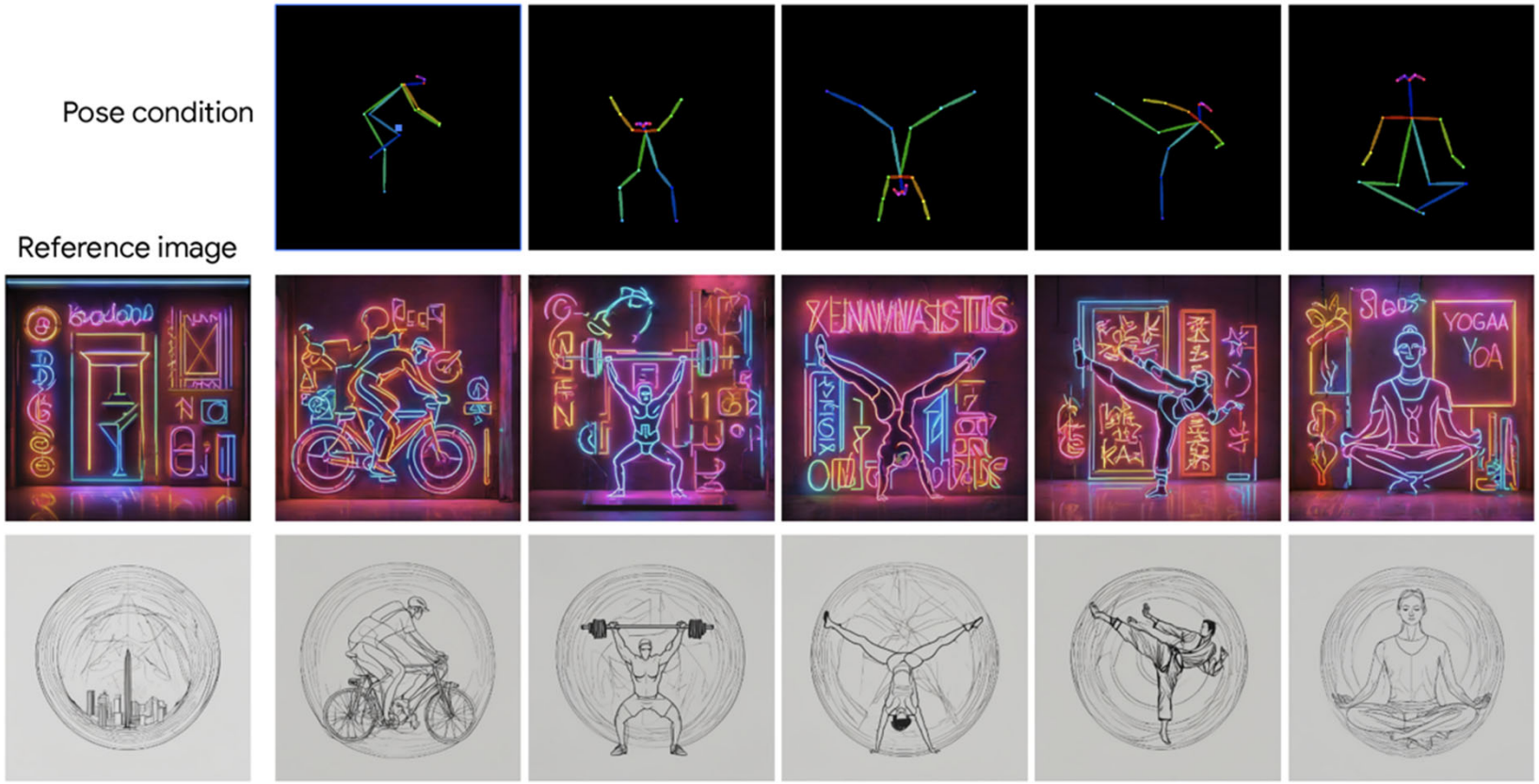
Style Aligned generation from an Input Image

StyleAligned with other methods

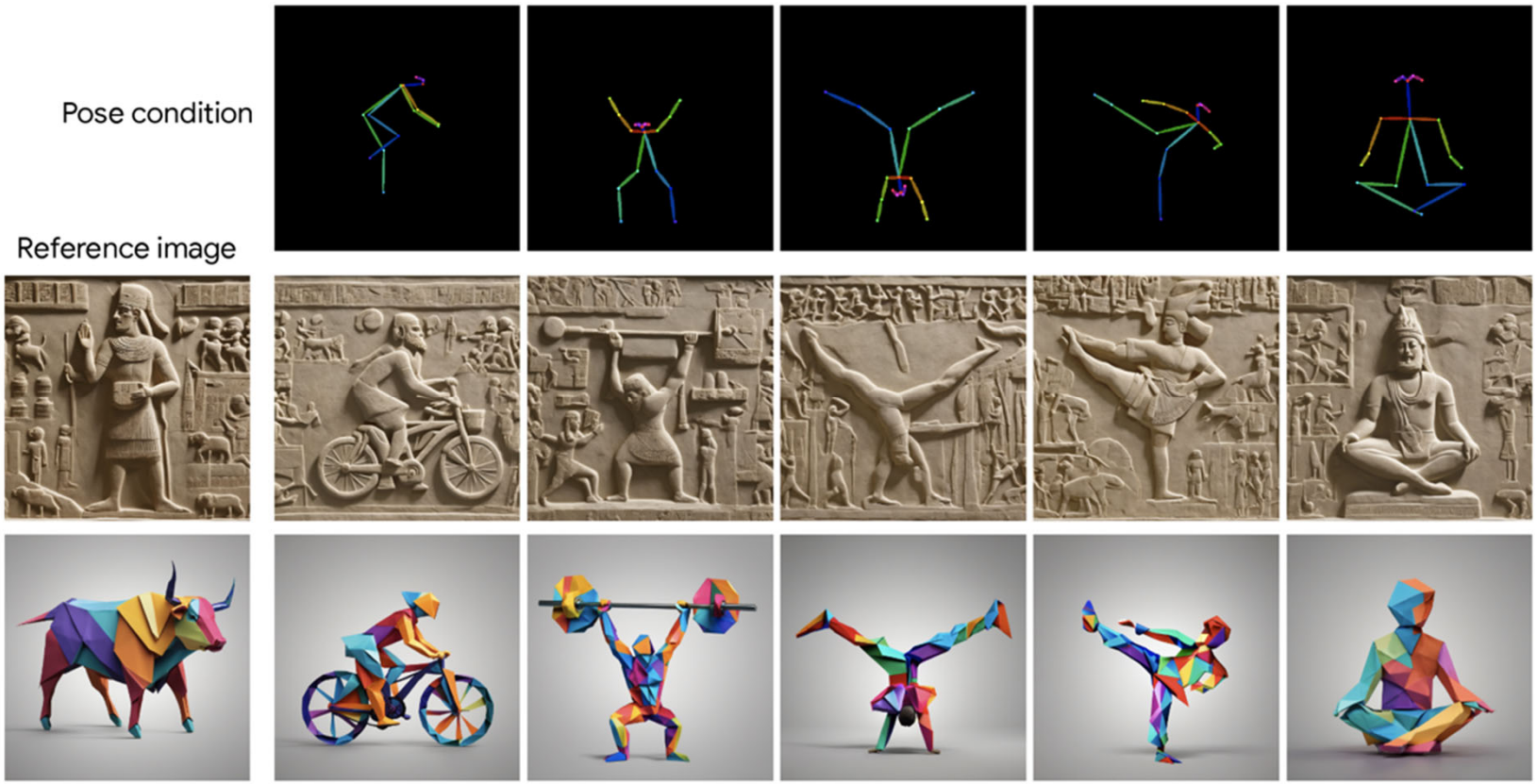
ControlNet + StyleAligned



ControlNet + StyleAligned

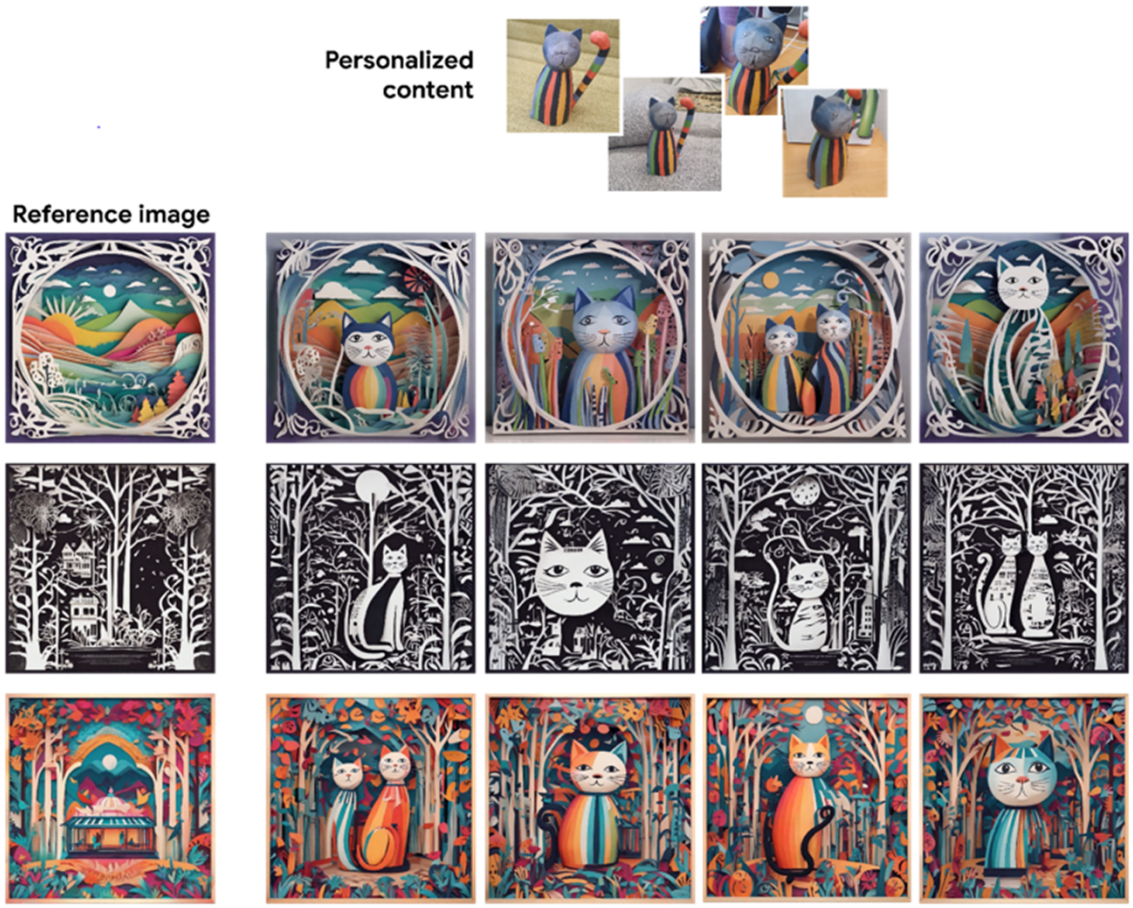


ControlNet + StyleAligned



Textual Inversion+Dreambooth

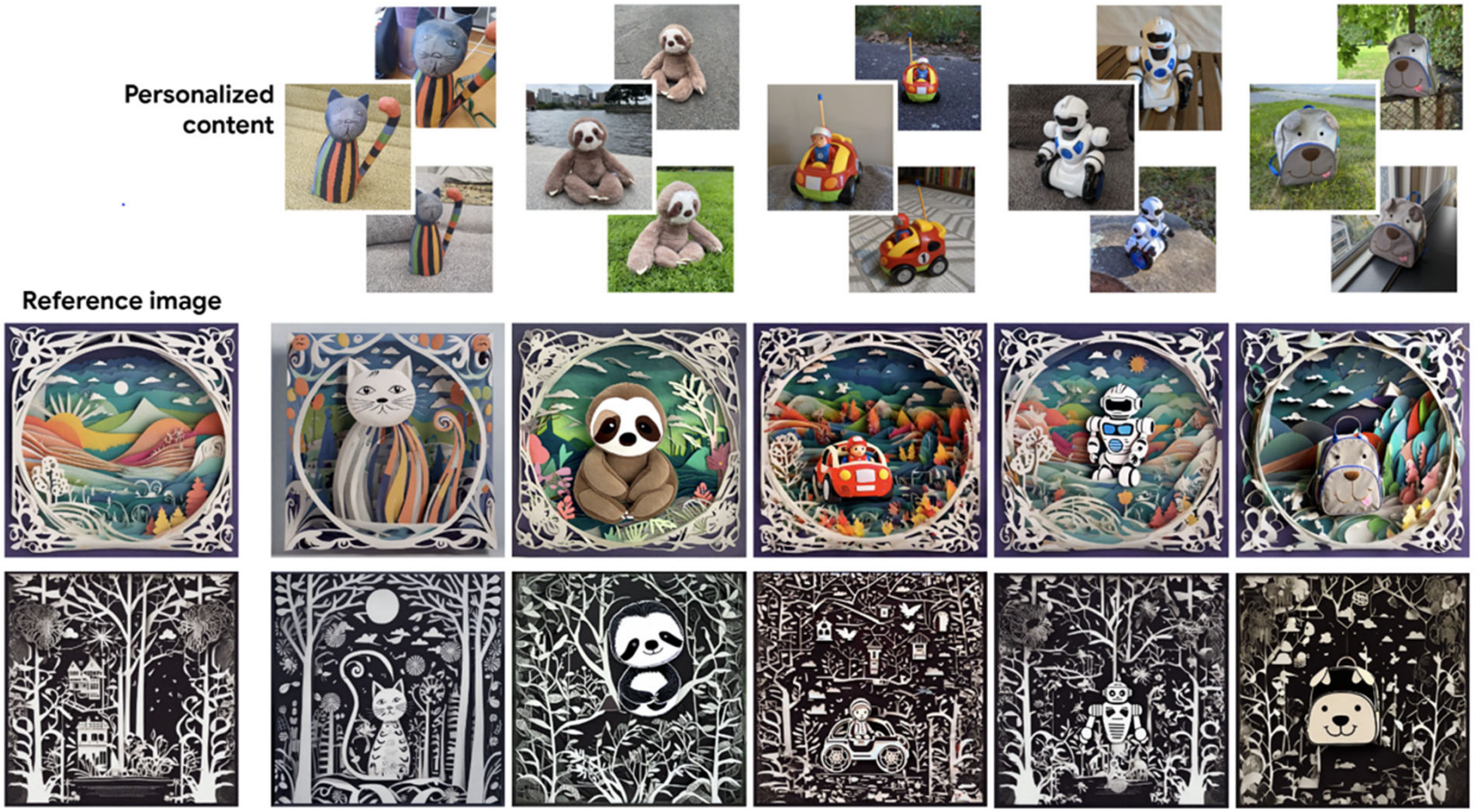
+ StyleAligned



W.O AdaIN


Personalized content
Reference image
"<V object> in the style of a beautiful paper-cut art."

DreamBooth + StyleAligned

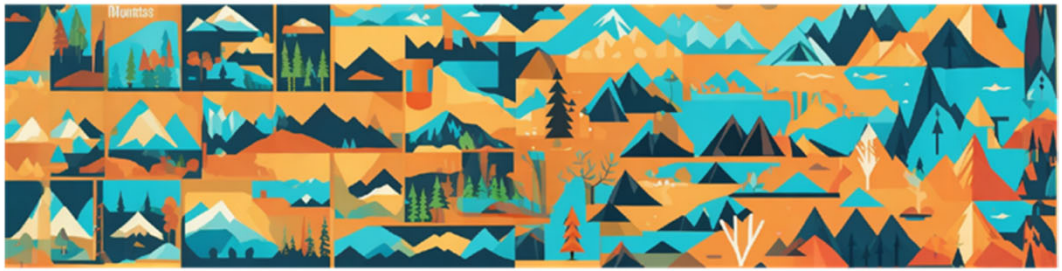


MultiDiffusion + StyleAligned

Reference image

"A poster in a flat design style."

"Houses in a flat design style."



"Mountains in a flat design style."

"Girrafes in a flat design style."

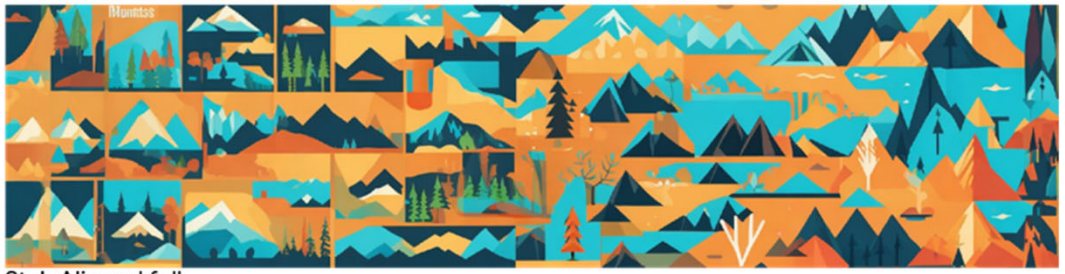
MultiDiffusion + StyleAligned

Reference image

"A poster in a flat design style."

W,O shared attention

W,O Attnetion AdaIN



StyleAligned full

MultiDiffusion in Multi Styles

Left Reference

Right Reference

MultiDiffusion in Multi Styles

Left reference

Right reference

Left Reference

Right Reference