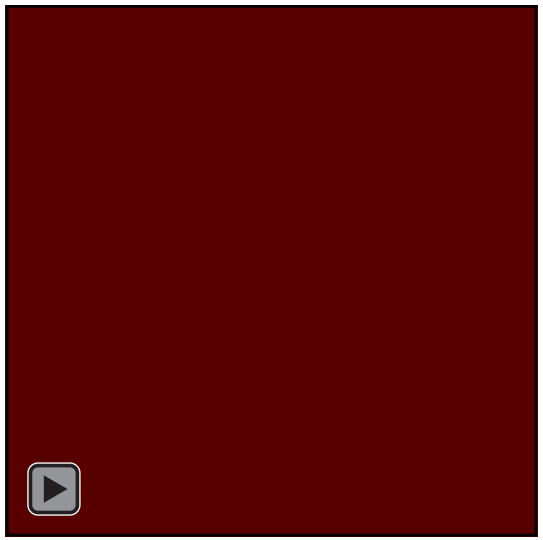
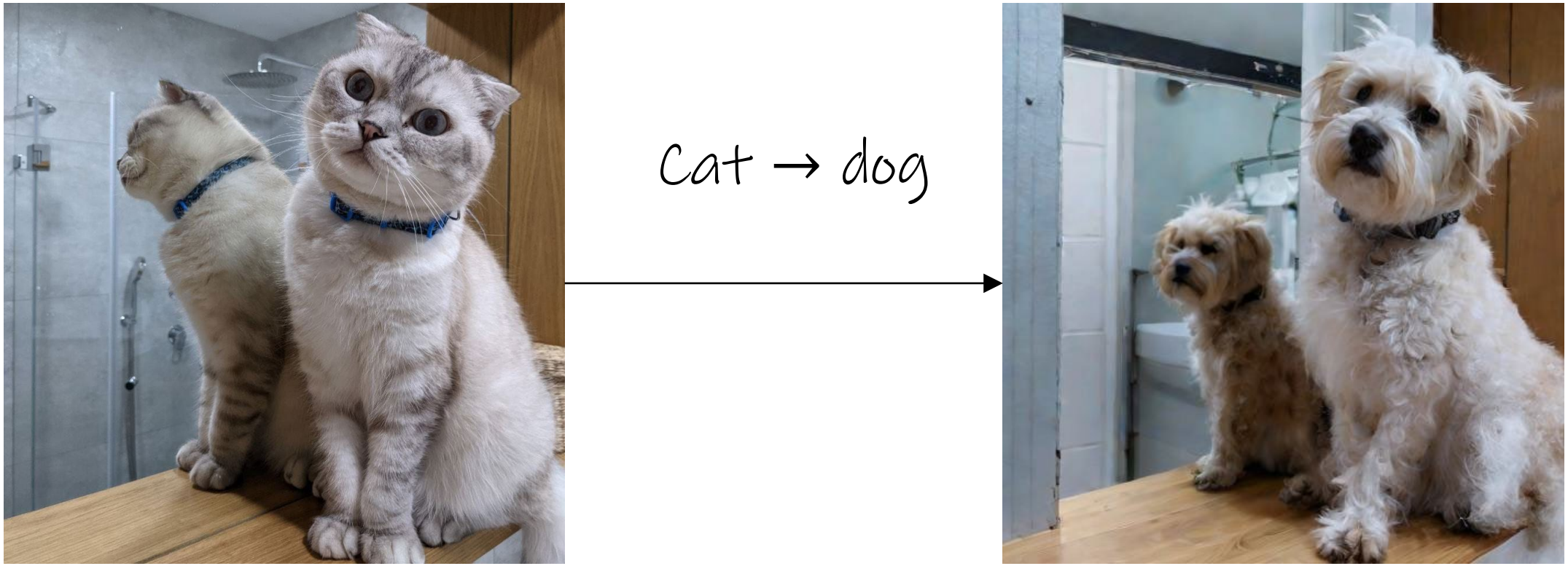
Attention in Large-scale Text-to-Image Models

Daniel Cohen-Or



TEL AVIV UNIVERSITY
אוניברסיטת תל אביב

Image Editing

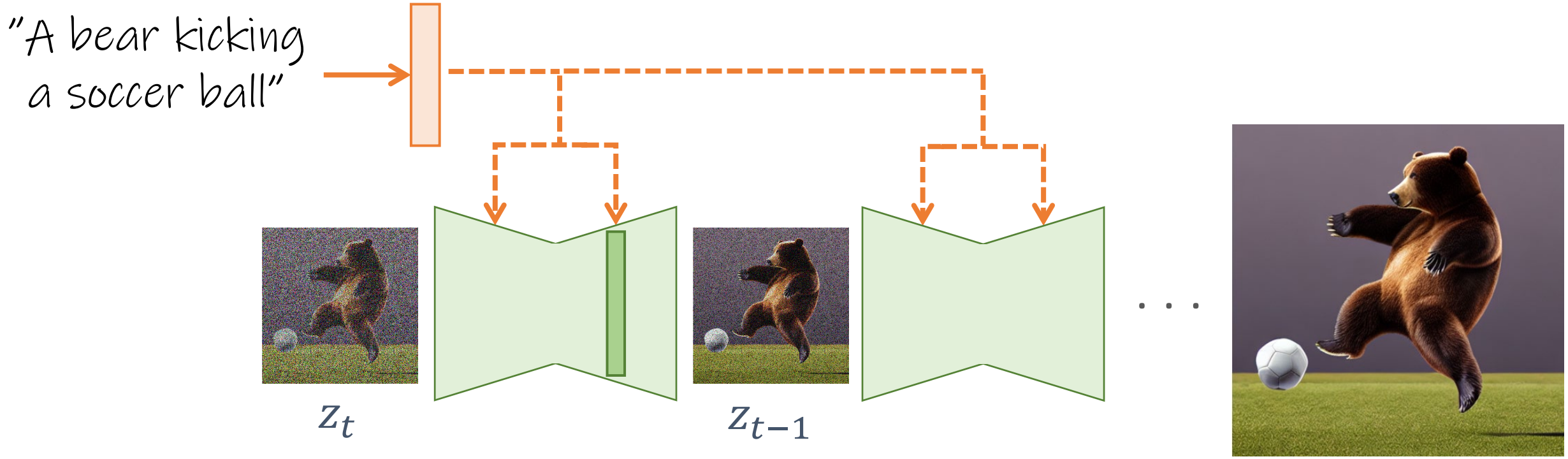


Text-Image Alignment

"A gray kitten and a ginger kitten and a black kitten and a white dog and a brown dog on a bed"

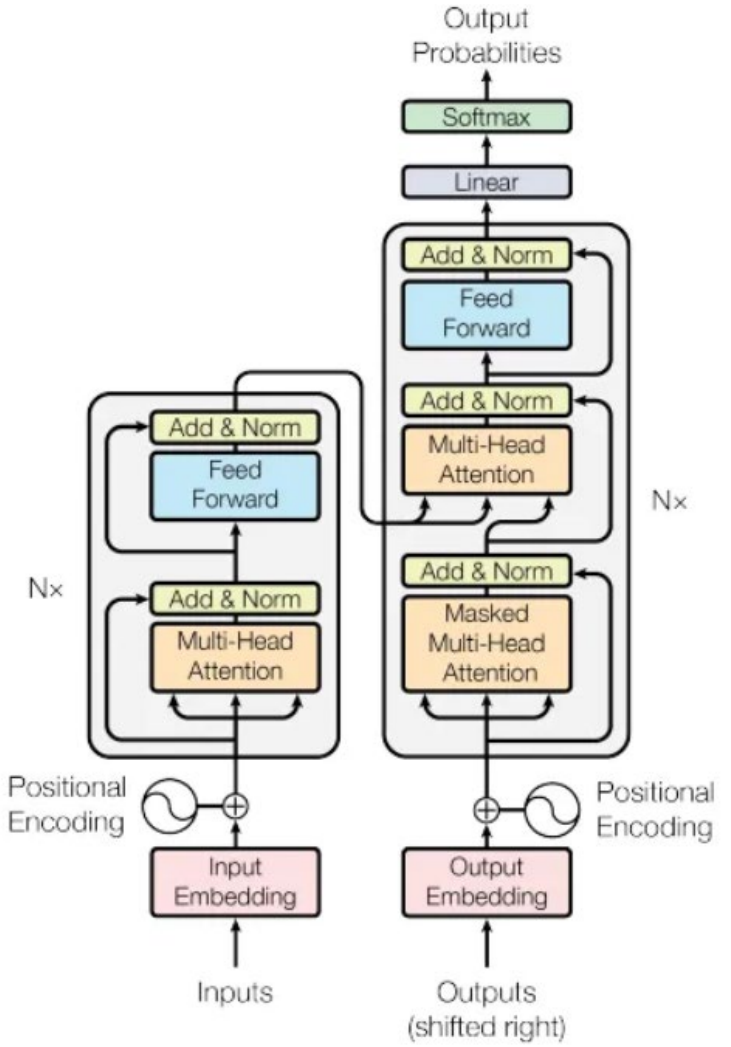
SDXL

The Denoising Process with Attention Layers



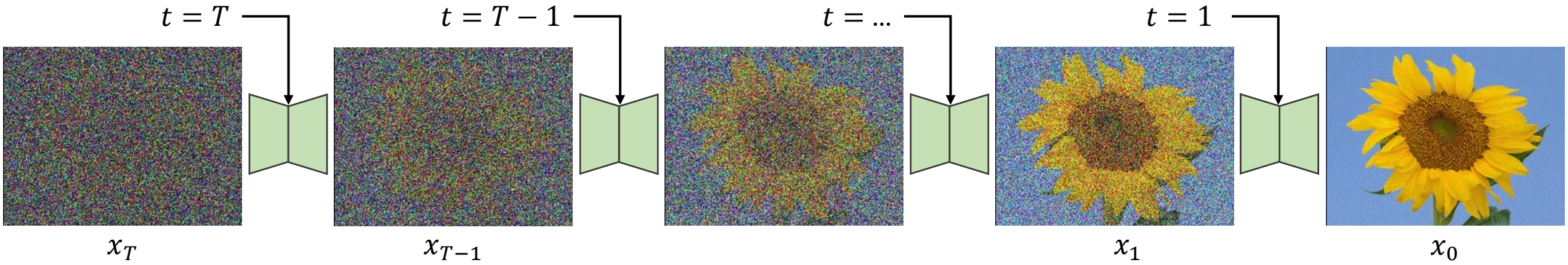
“Attention is all you need” , Waswani et al. 2007

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

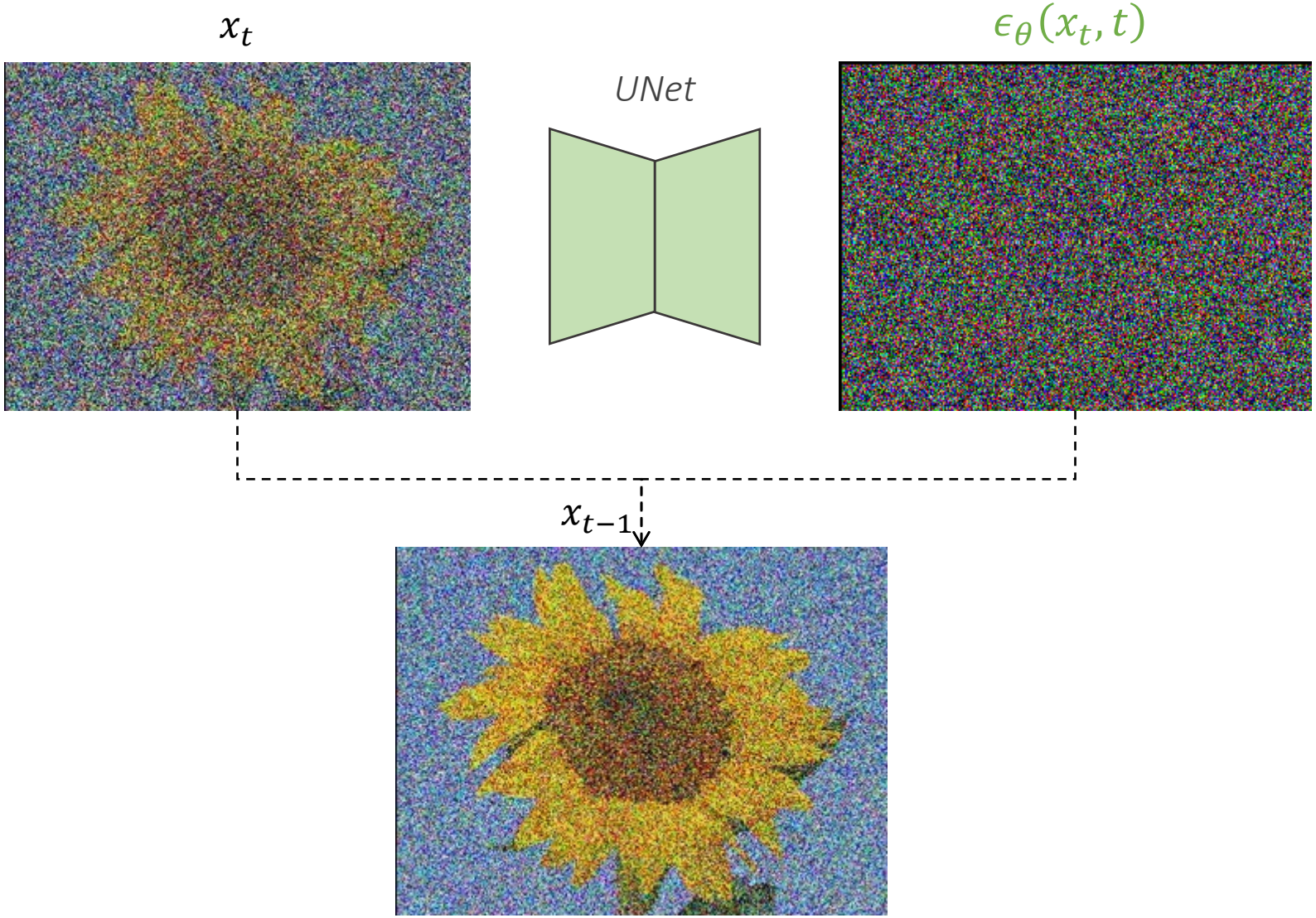


From “Attention is all you need” paper by Vaswani, et al., 2017 [1]

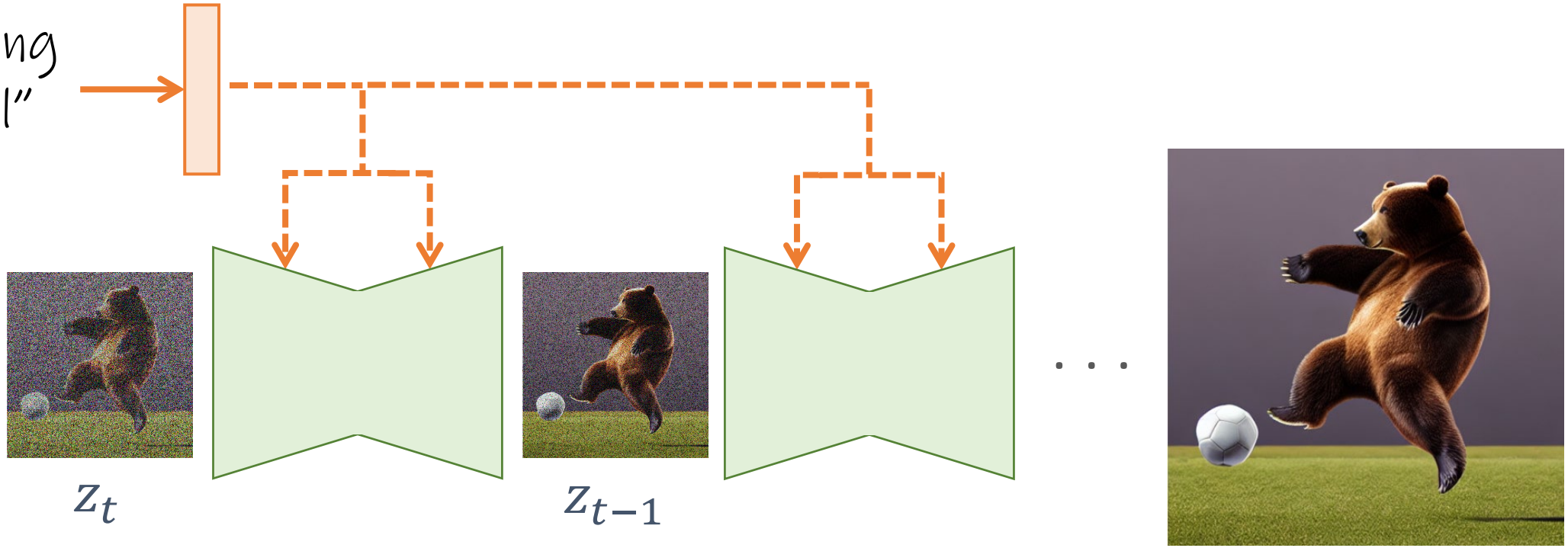
Diffusion Models

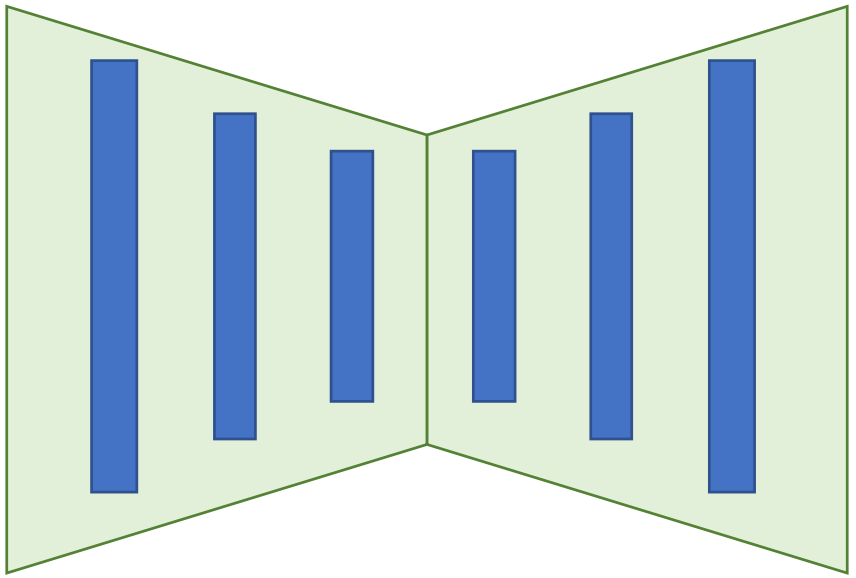


Diffusion Models

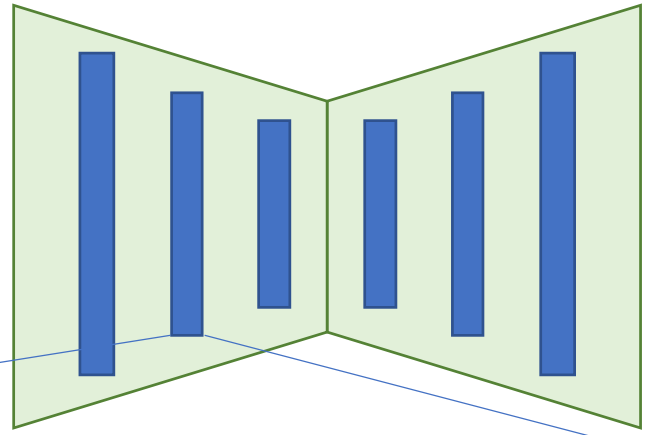


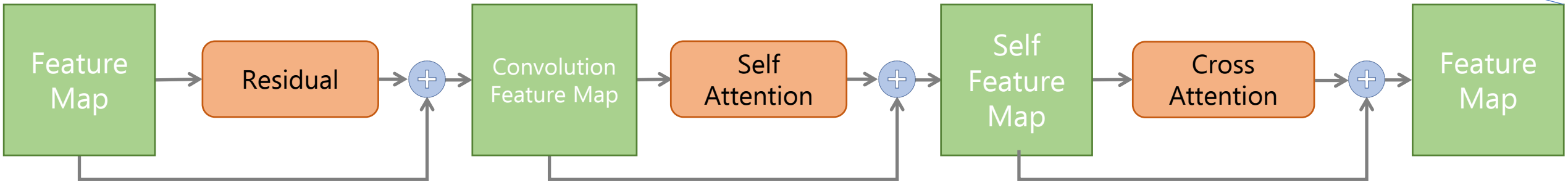
"A bear kicking a soccer ball"





The UNet Layer





Cross-Attention

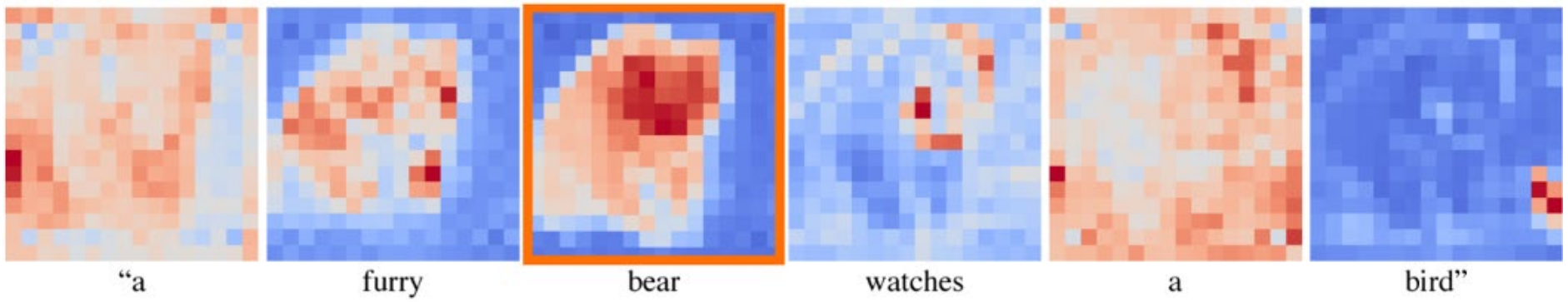
$N = 6$

P^2

A
furry
bear
watches
a
bird

Q K

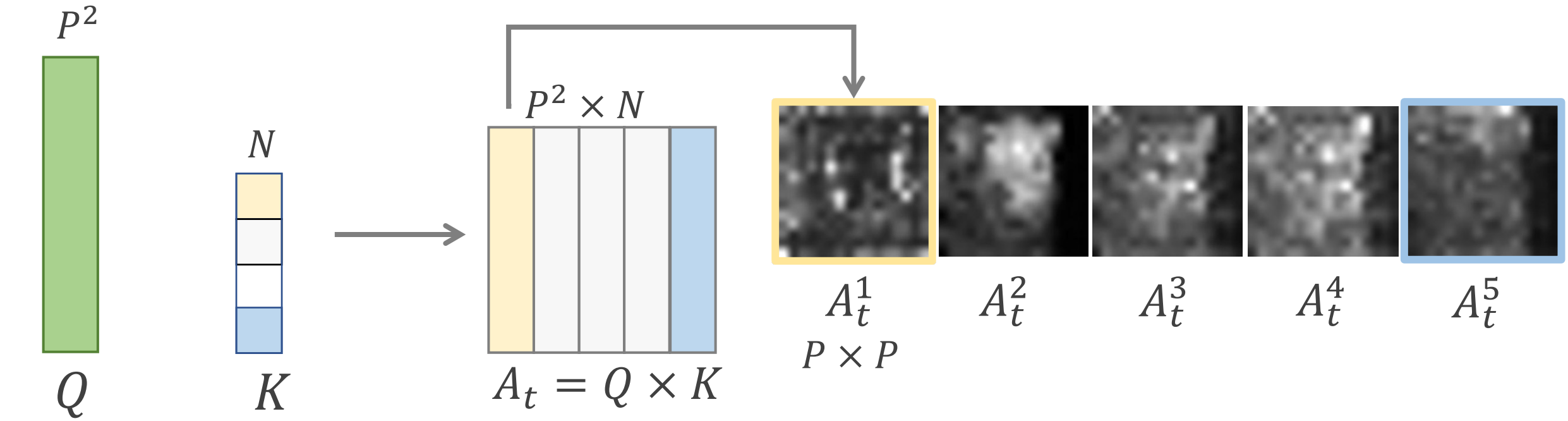
$P^2 \times N$



$$A_t = Q \times K$$

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

Cross-Attention



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

Cross-attention

Query

X

X

Key

“Bird”

“Bear”

Value

Cross-attention

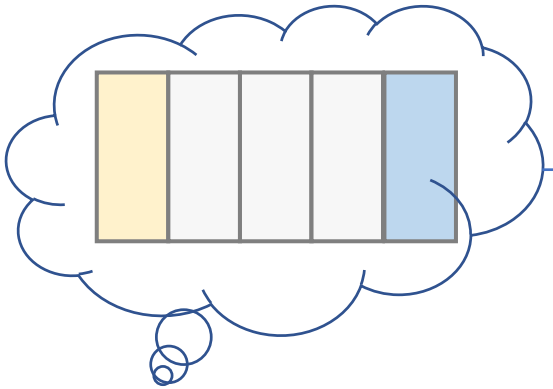
Query

QK^T

$(QK^T)V$

X

X



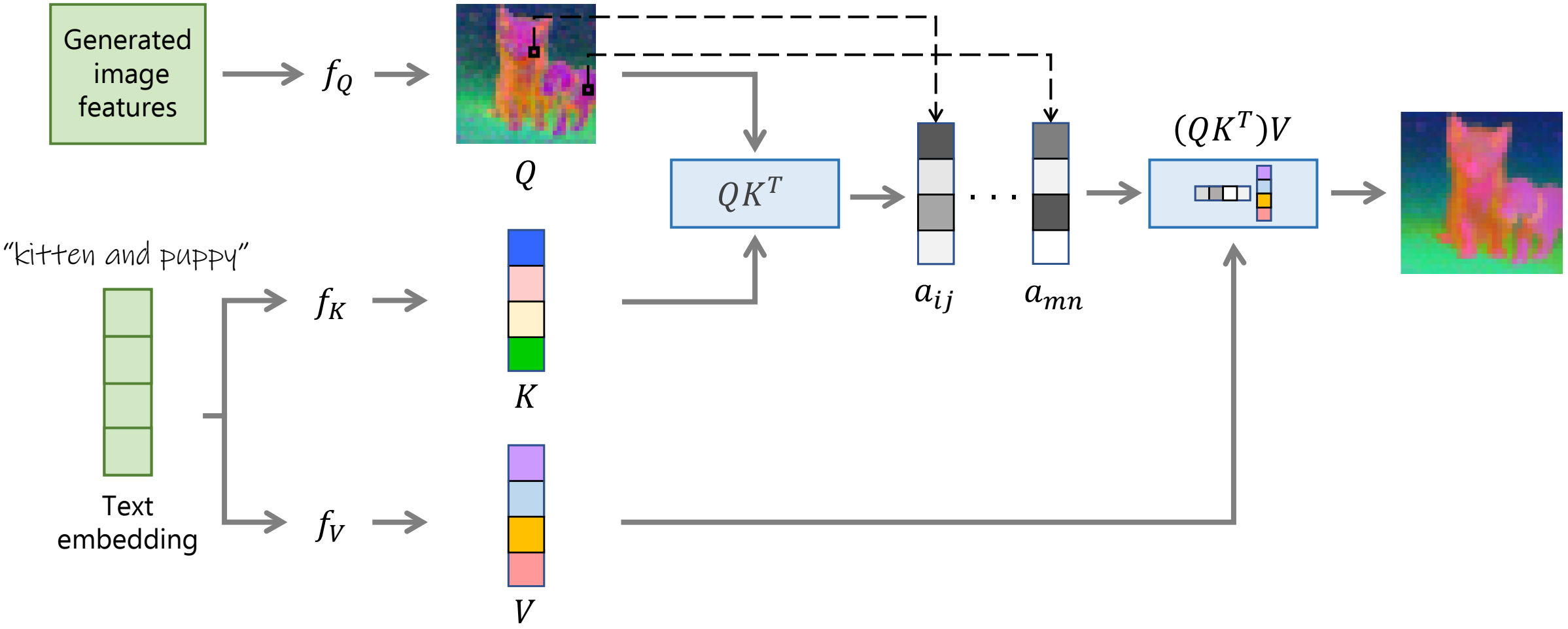
Key

"Bird"

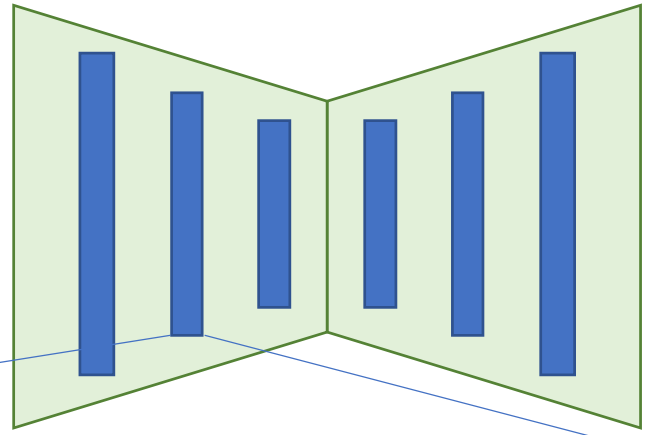
"Bear"

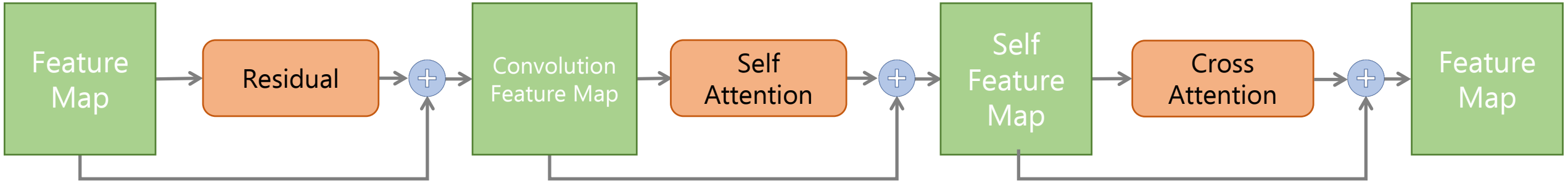
Value

Cross-Attention Layers



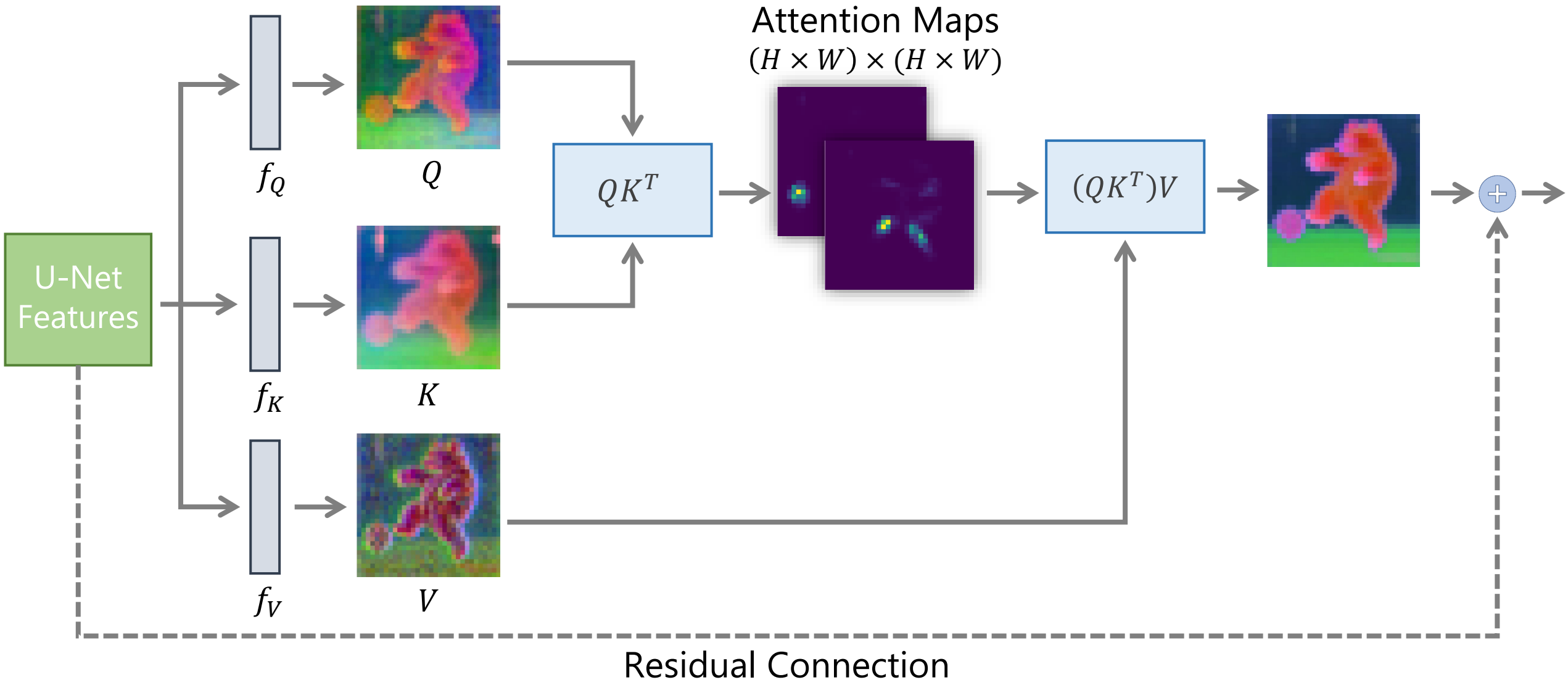
The UNet Layer



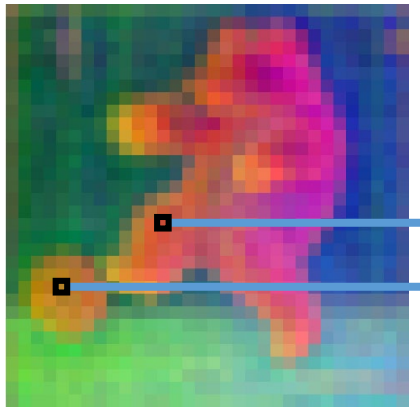


The Self-Attention

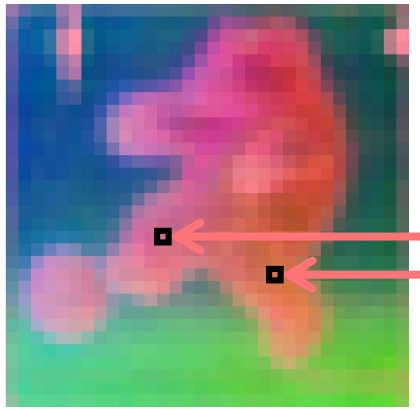
Represents where the model "looks" in the image for each spatial position in Q



The Self-Attention

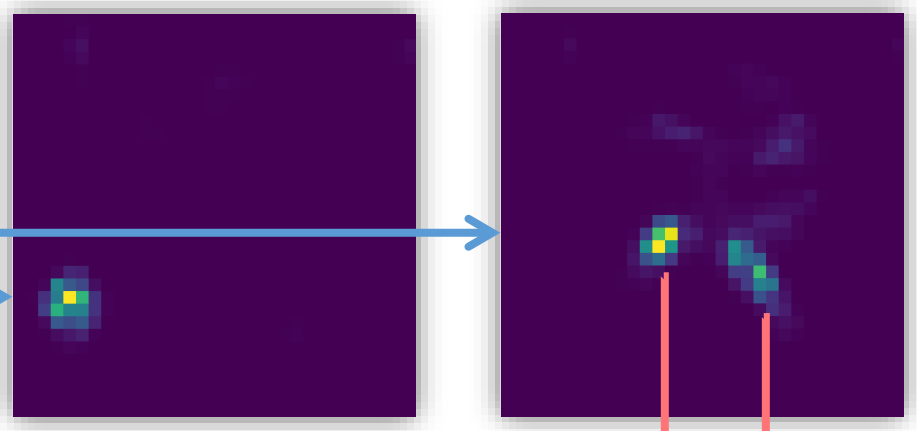


Queries
$H \times W \times C$



Keys
$H \times W \times C$

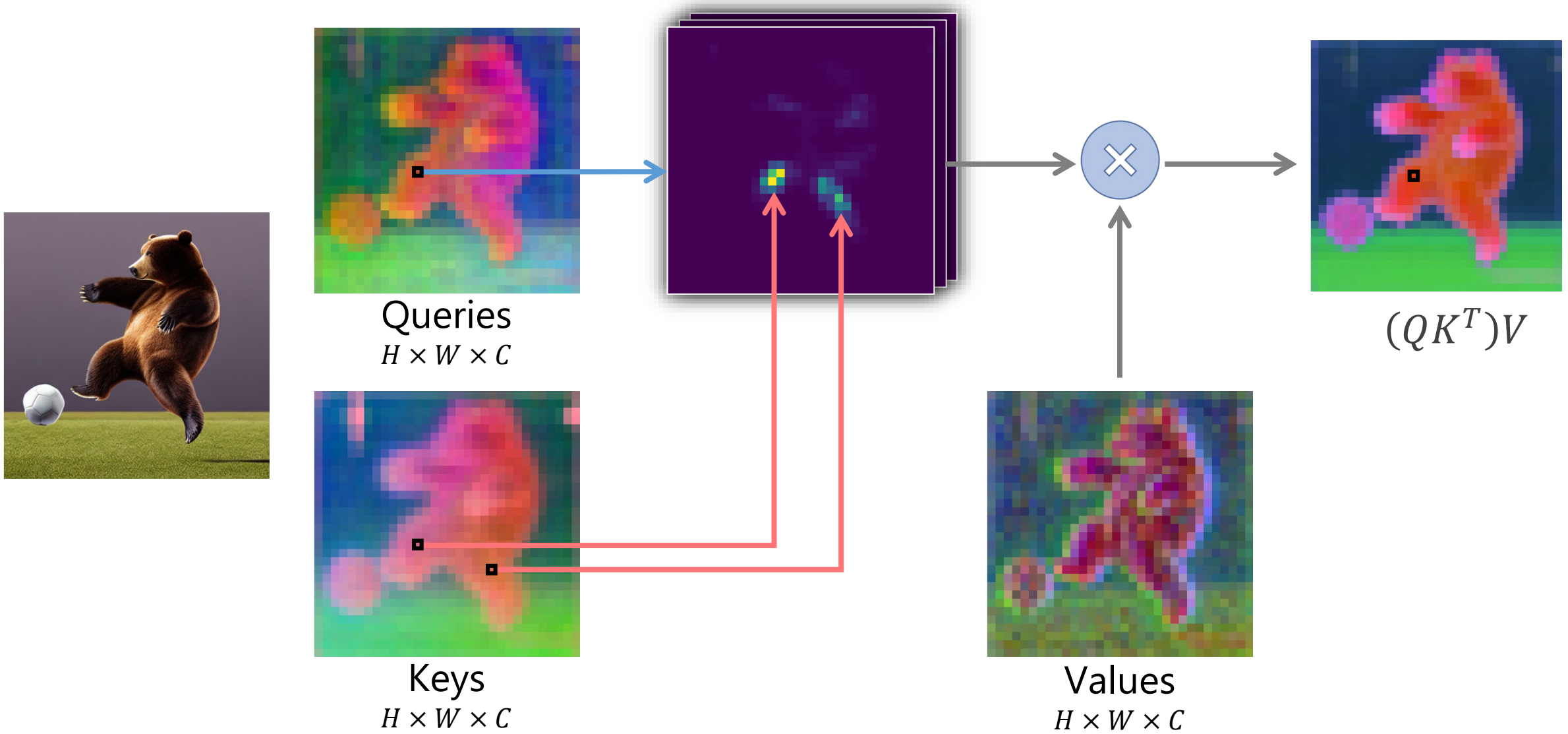
Attention Maps
$(H \times W) \times (H \times W)$



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



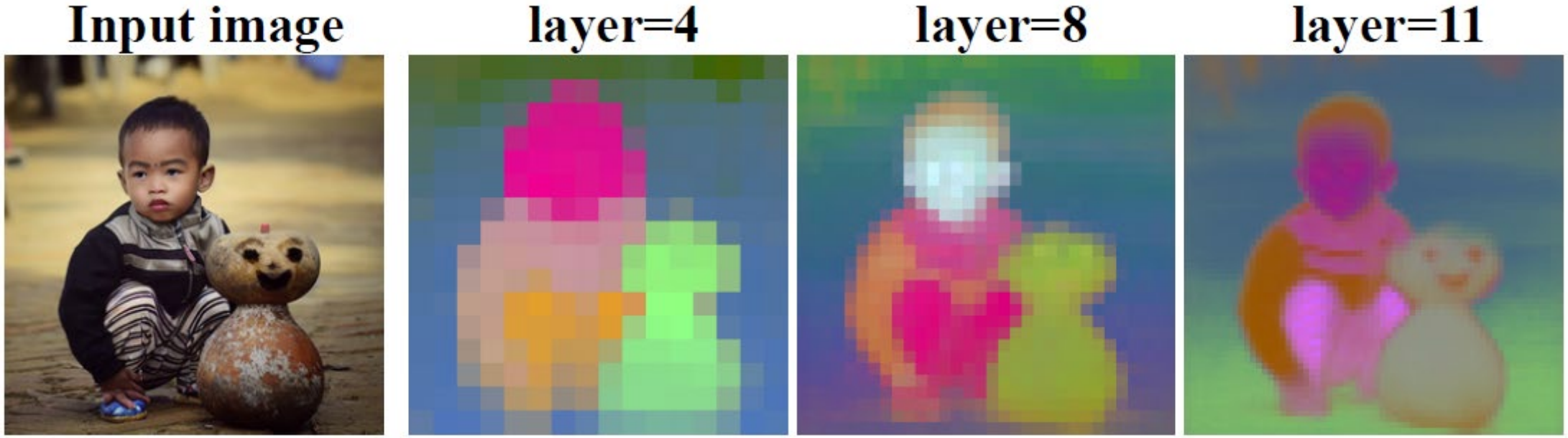
t = 0.6, layer: 10 / 70

t = 0.6, layer: 35 / 70

Self-Segmentation

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

Self-Attention Maps



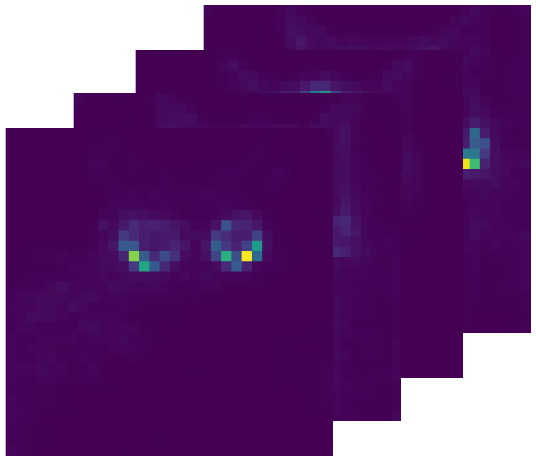
Are these PCA on the self-attention ? On what exactly the QK maps?

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

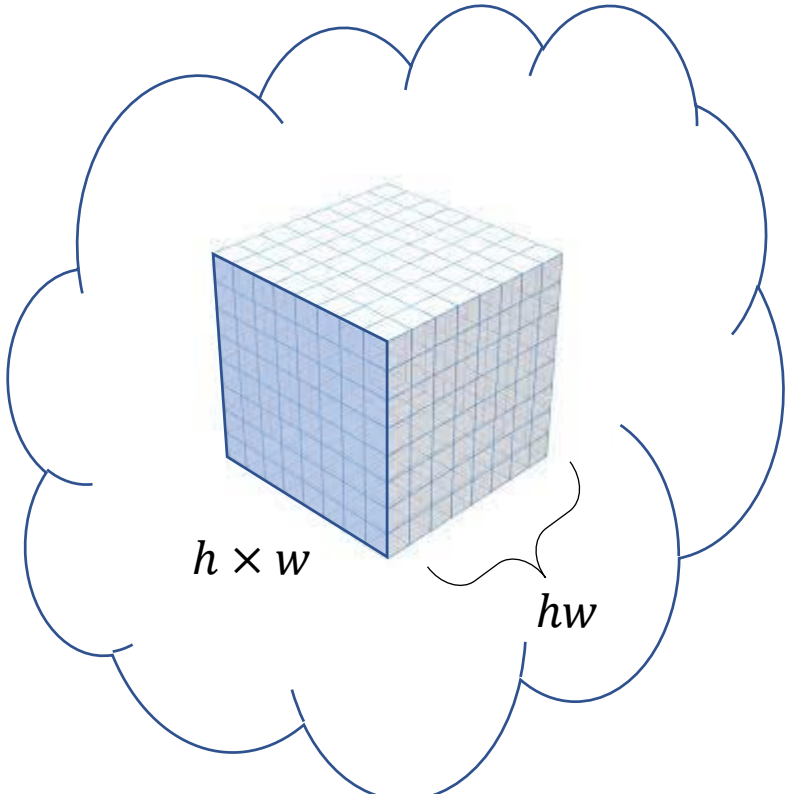
Self-Segmentation

"a **cat** is wearing sunglasses"

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



$hw \times (h \times w)$

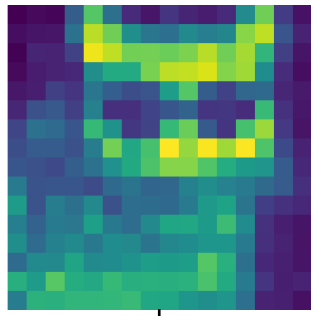


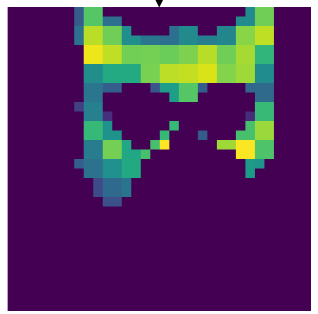
cluster

Segments labeling

“a cat is wearing sunglasses”

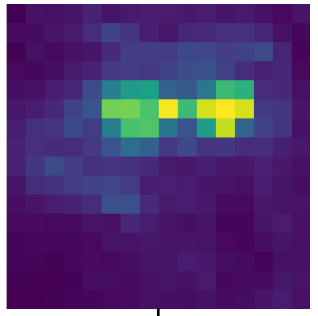
cat

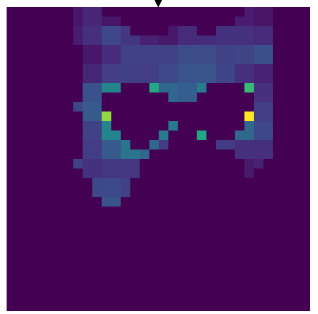




score: 0.65

sunglasses

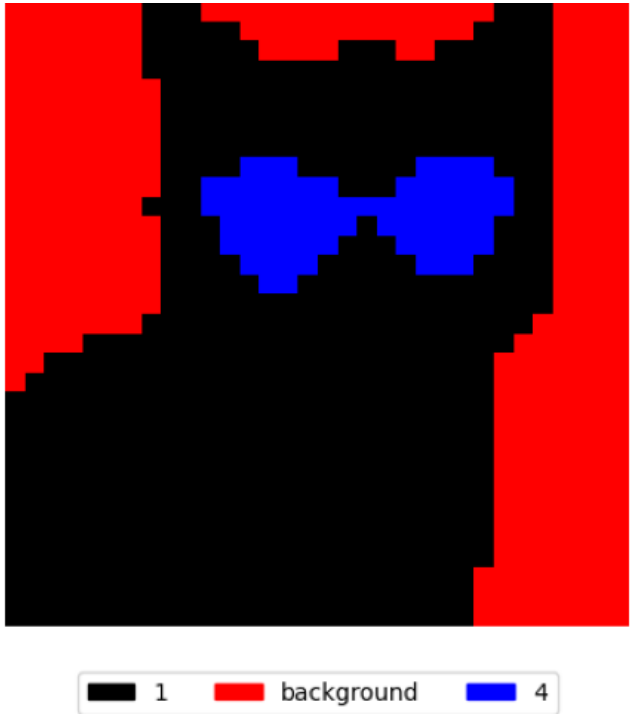




score: 0.19

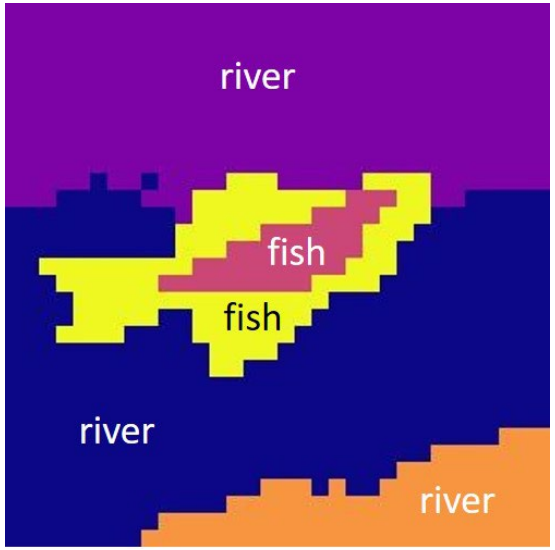
Segments labeling

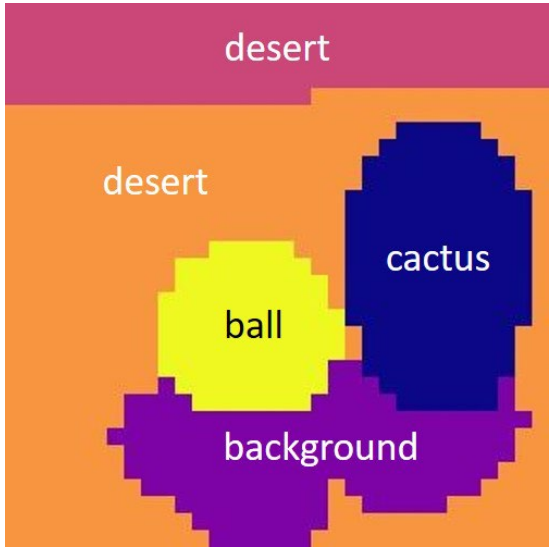
“a cat is wearing sunglasses”

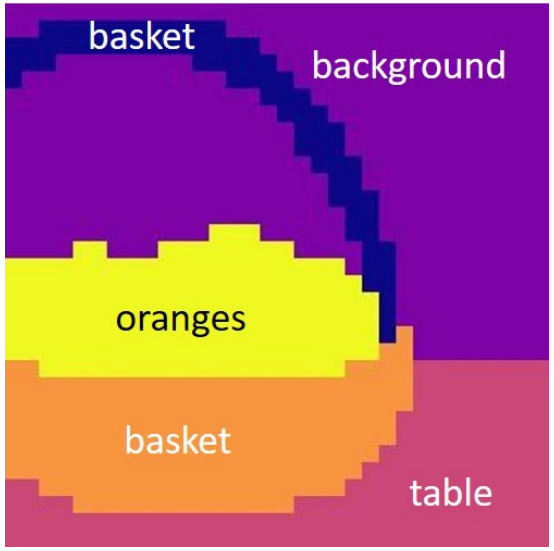


1-cat, 4-sunglasses

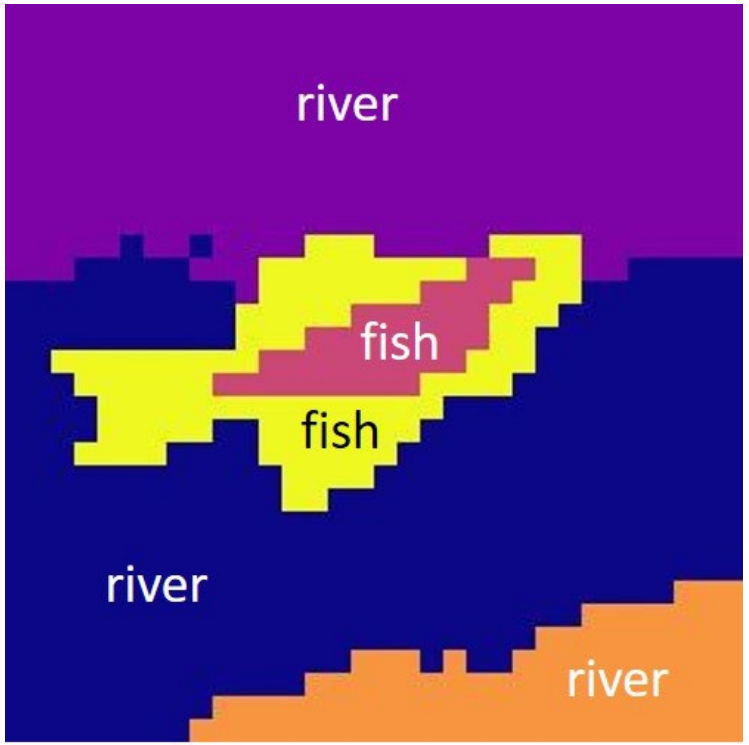
Self-Segmentation Results







Self-Segmentation Results



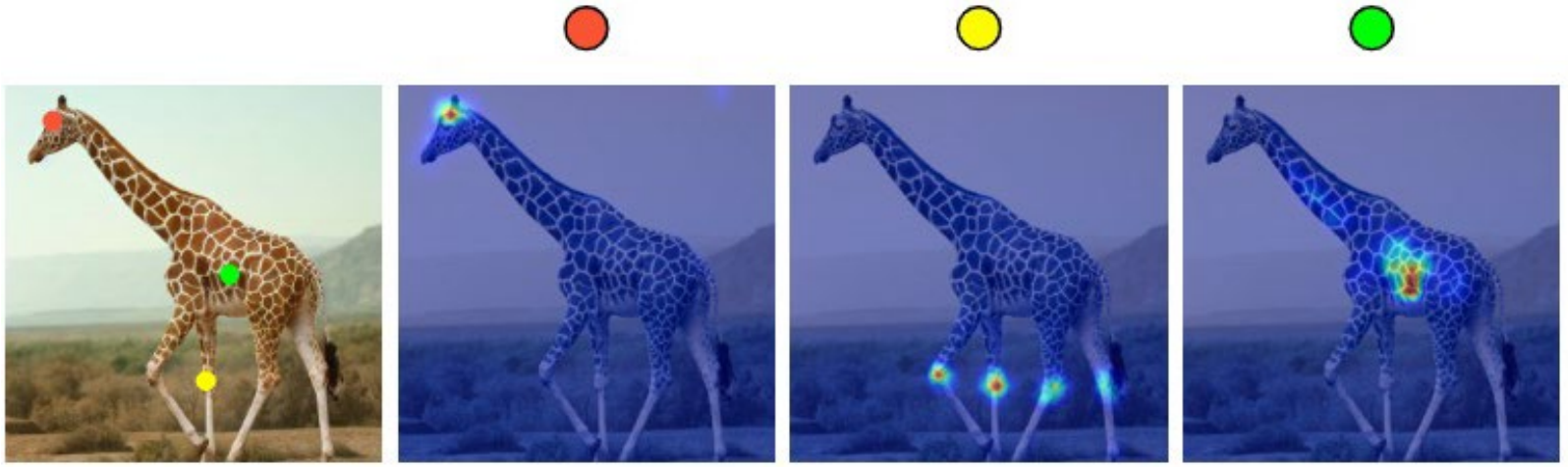
Cross-Image Attention

Structure

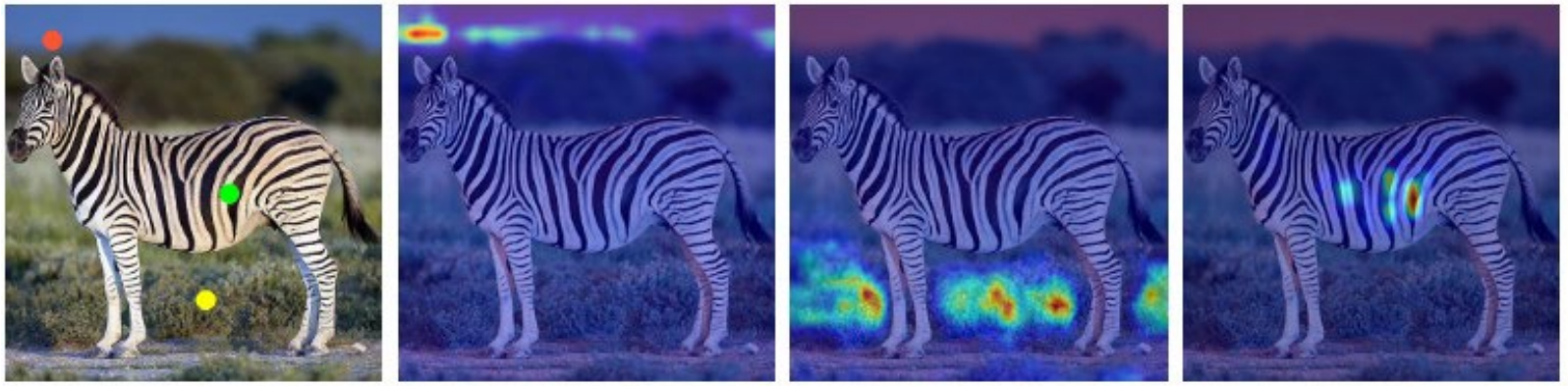
Appearance

Output

The Roles of the Queries, Keys, and Values



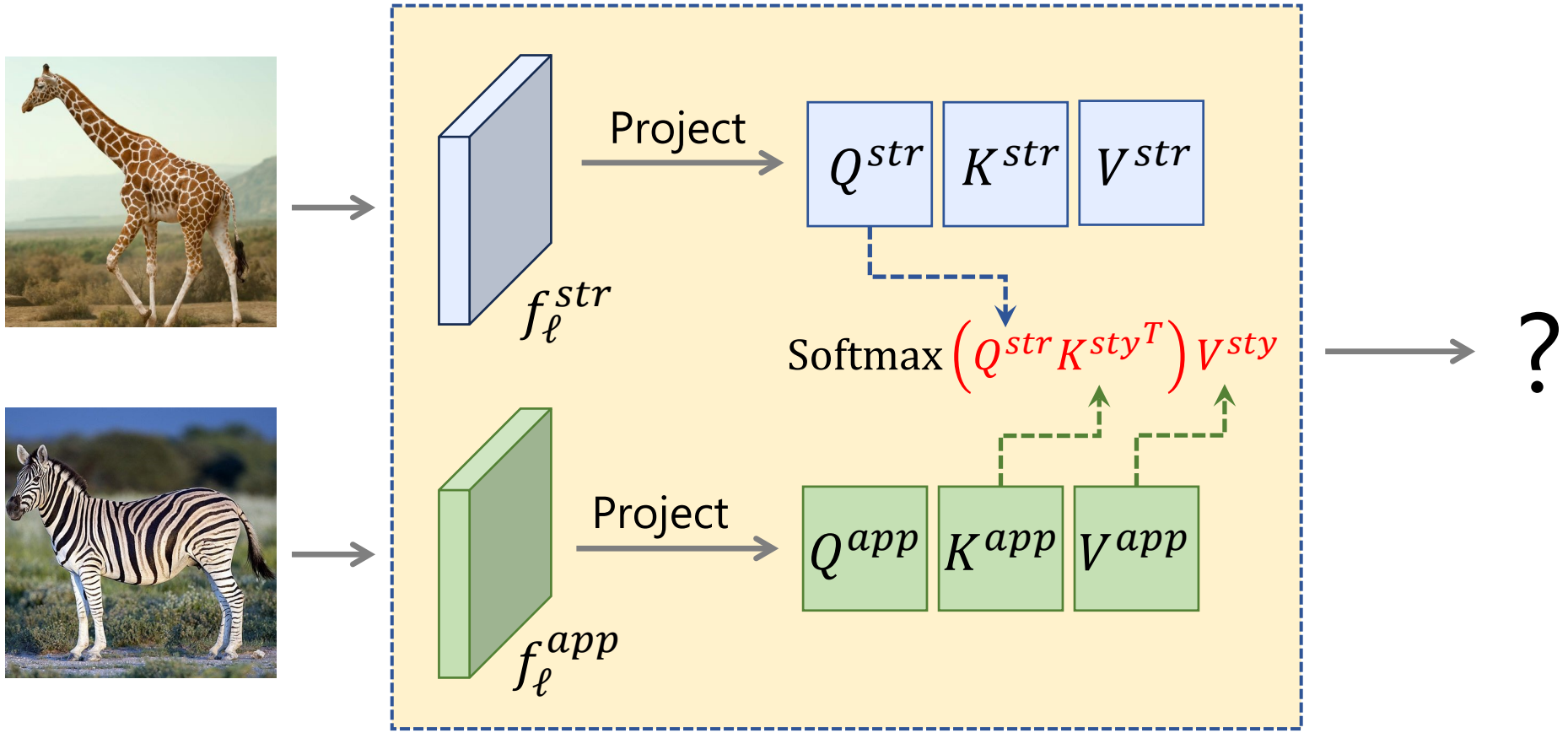
$$Q_{struct} \cdot K_{struct}^T$$



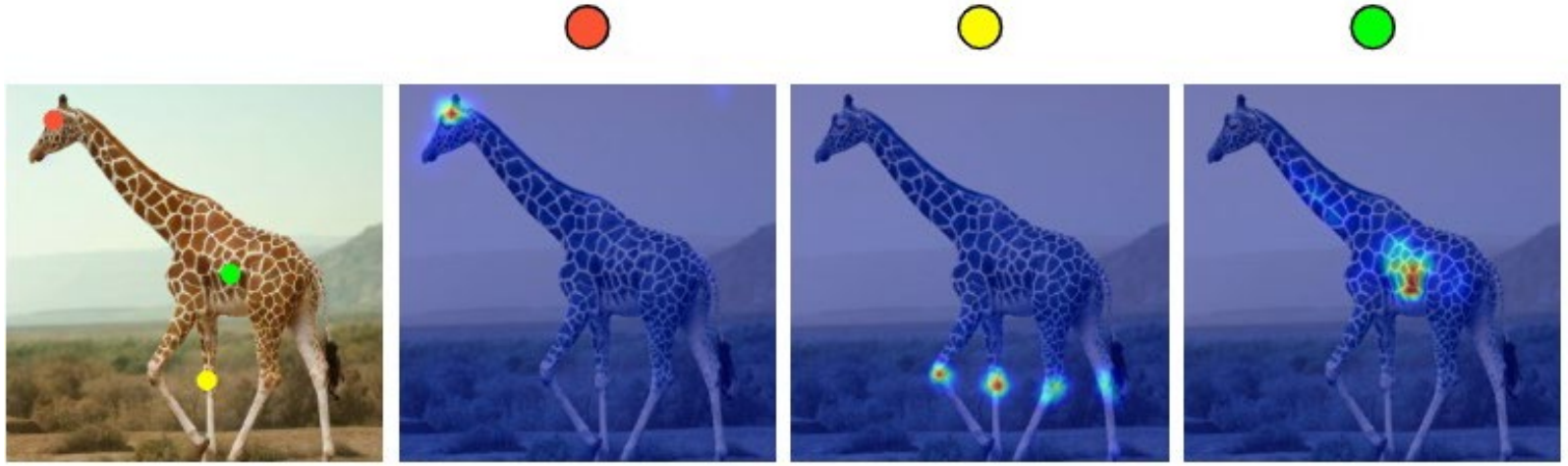
$$Q_{app} \cdot K_{app}^T$$

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

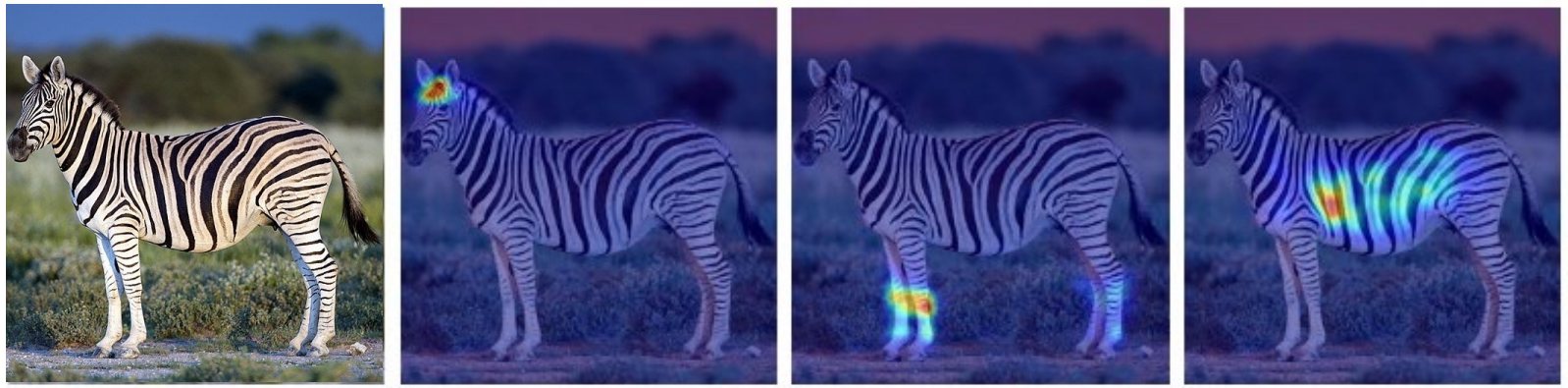
What If we Swapped the Queries, Keys, and Values Between Different Images?



The Roles of the Queries, Keys, and Values



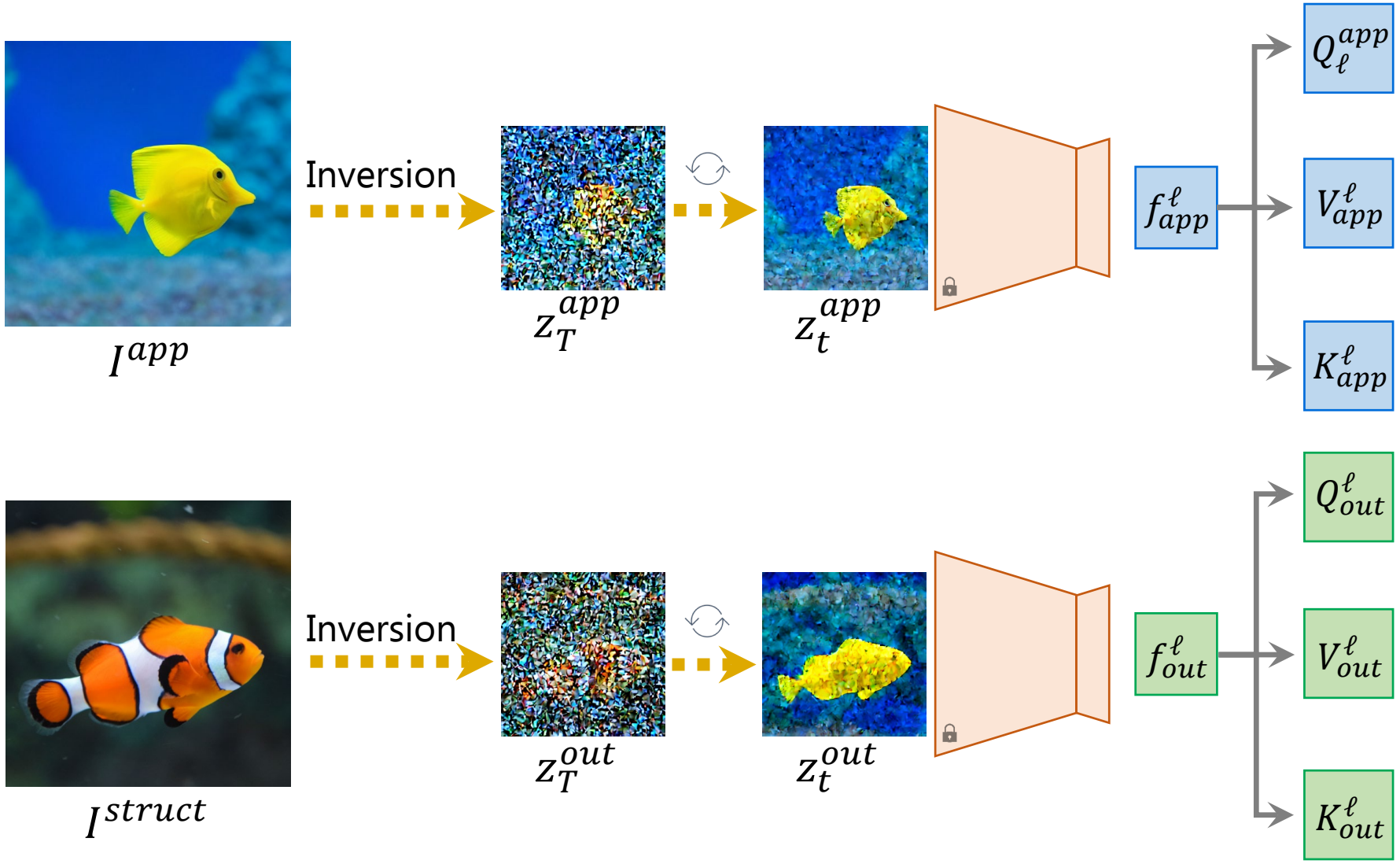
$$Q_{struct} \cdot K_{struct}^T$$



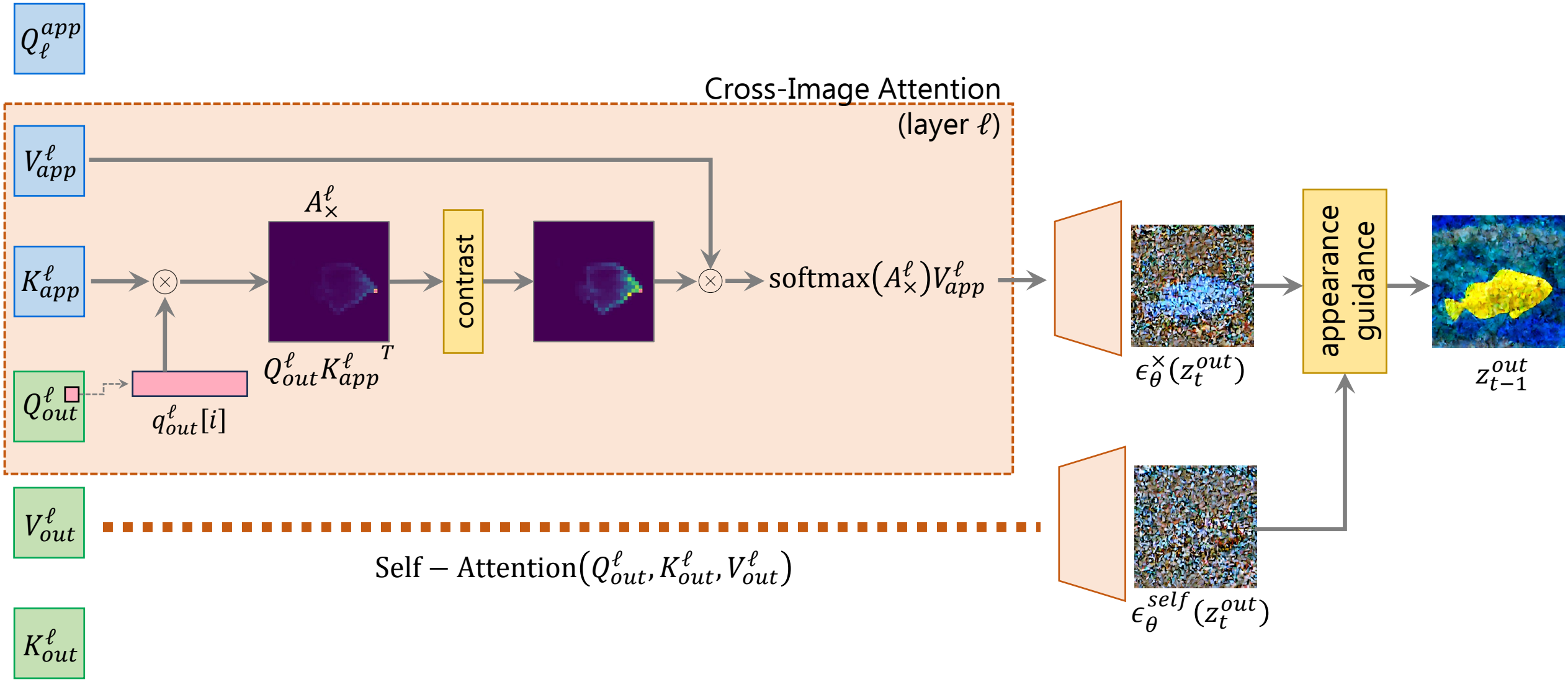
$$Q_{struct} \cdot K_{app}^T$$

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

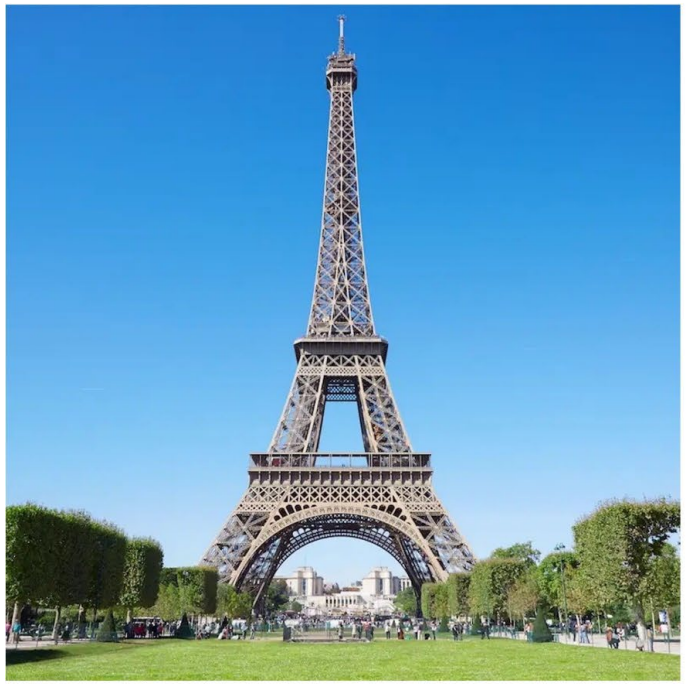
Structure

Appearance

Output

Appearance Transfer Results

Eiffel Tower



Structure

Sagrada Família

Appearance

Output

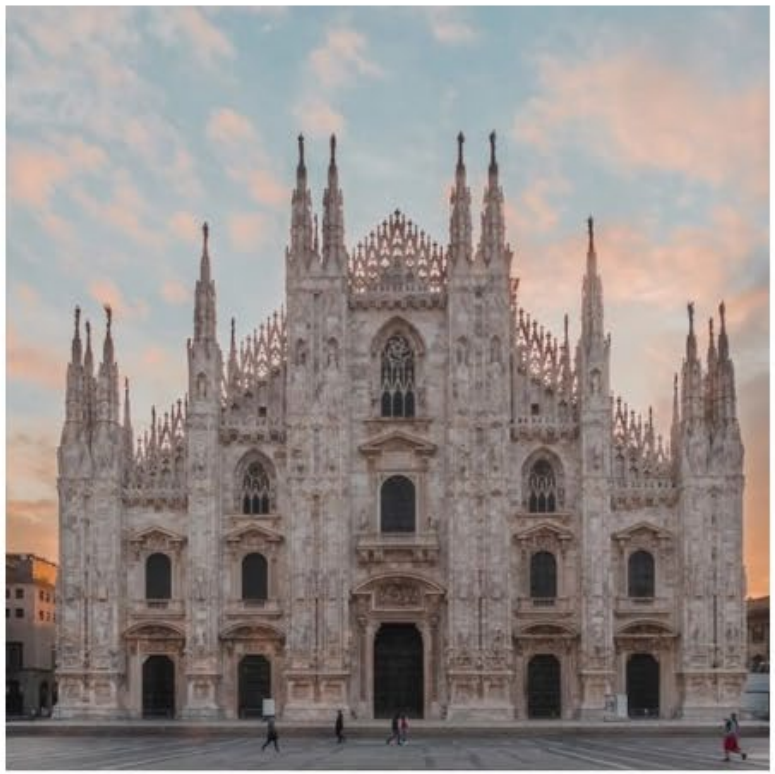
Appearance Transfer Results

Structure

Appearance

Output

Appearance Transfer Results



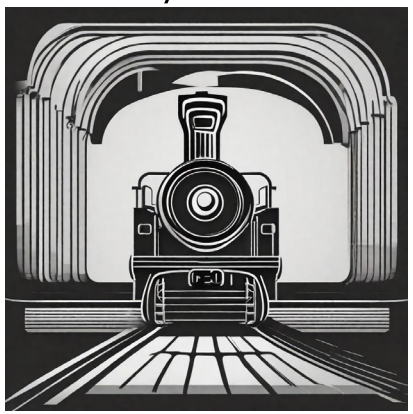
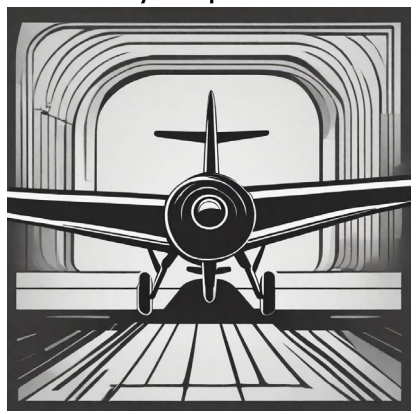
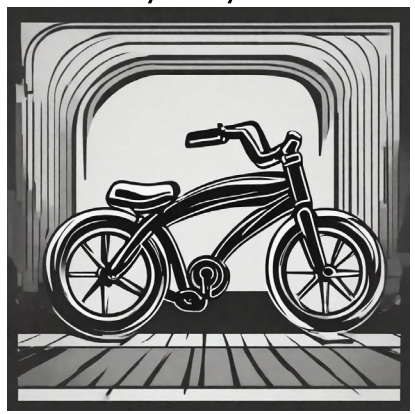
Structure

Appearance

Output

StyleAligned

(link)

“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..."

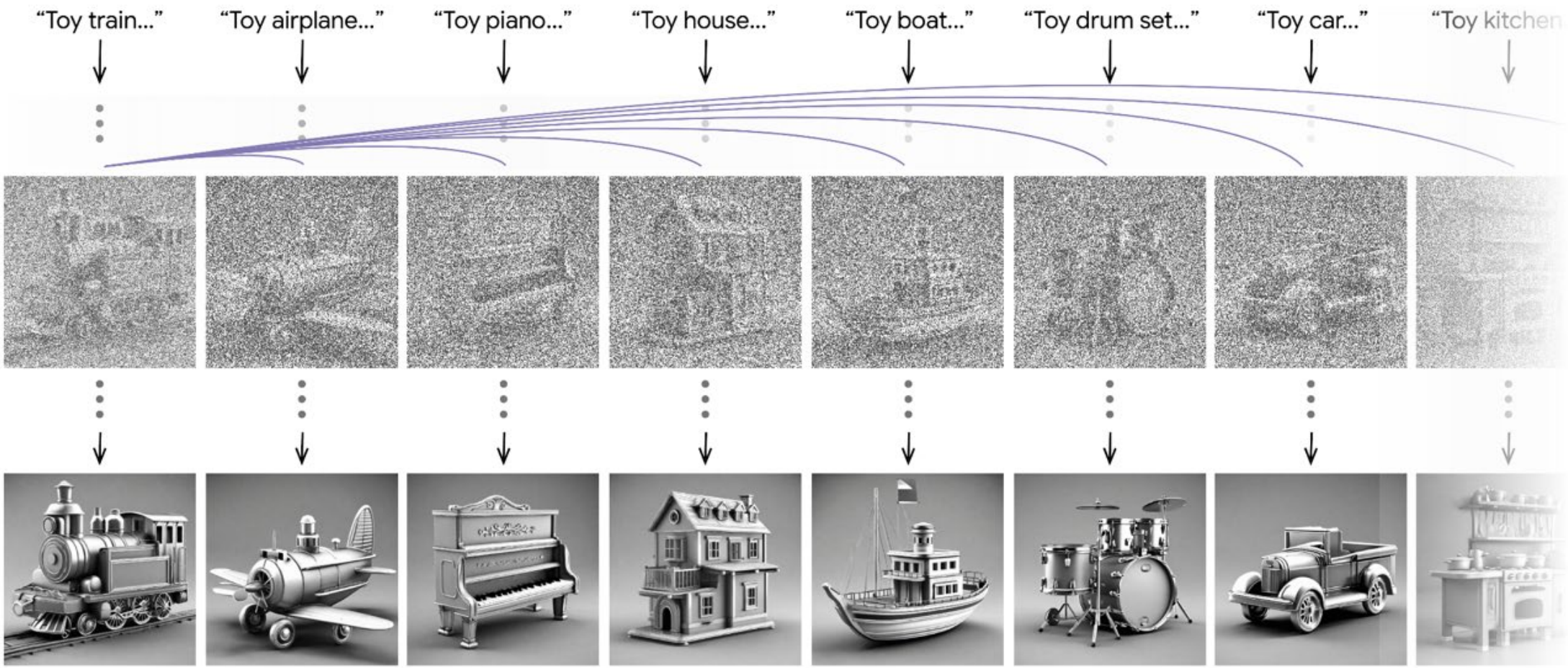
"A person riding a bike..."

"A cactus..."

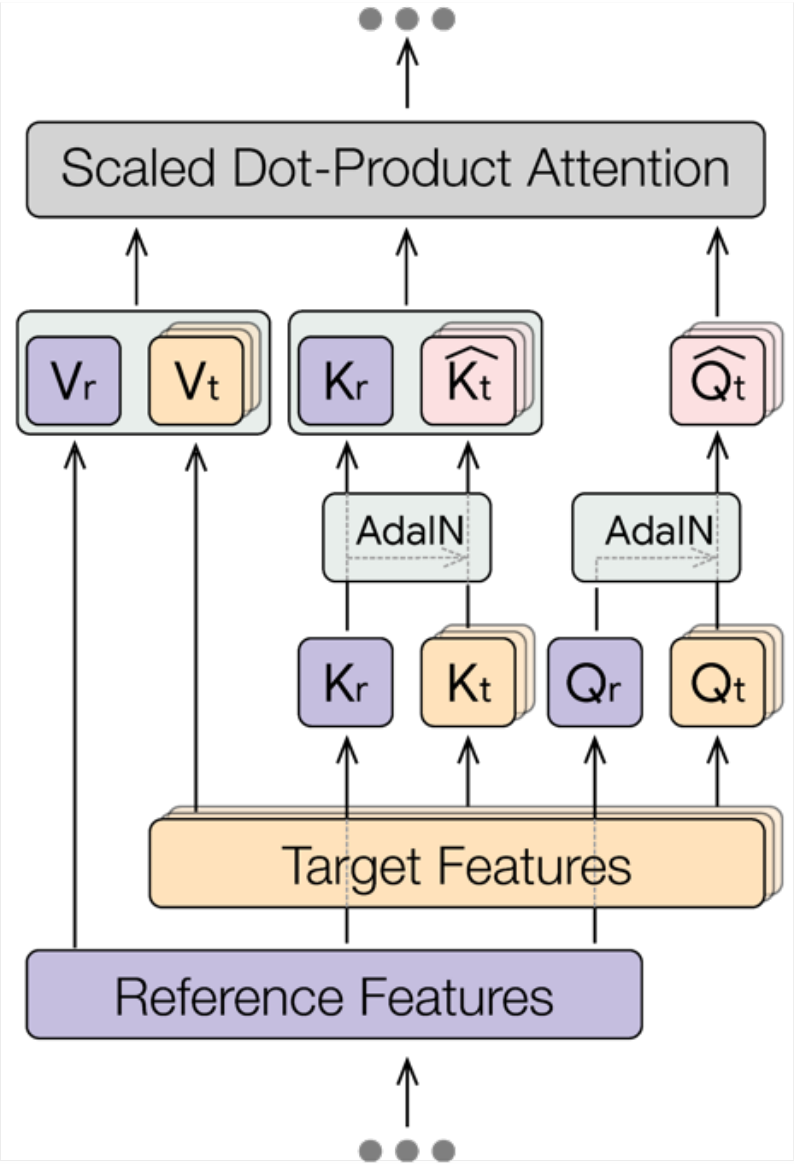
"... in minimal origami style."

Text-to-Image Generation with StyleAligned

Shared attention during the diffusion process



Shared Attention Layer



Style Aligned generation of Synthetic Images

“Firewoman...”

“ Gardner...”

“ Scientist ...”

“ Police woman...”

“ Saxophone player...”

“Painter...”

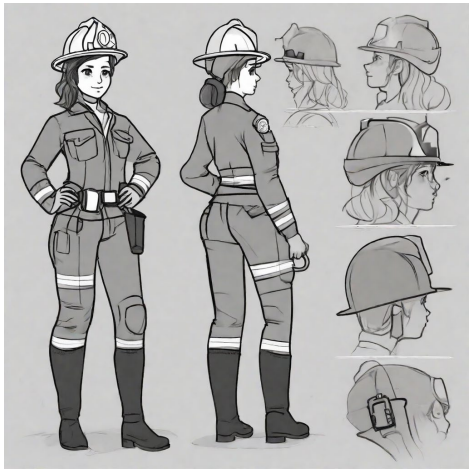
“Astronaut...”

“Taxi Driver...”

“ ...made of claymation, stop motion animation.”

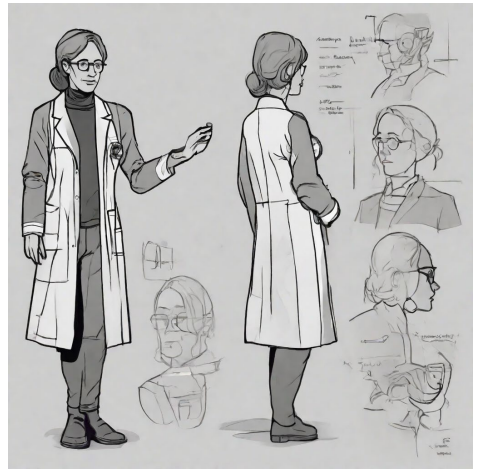
Style Aligned generation of Synthetic Images

“Firewoman...”

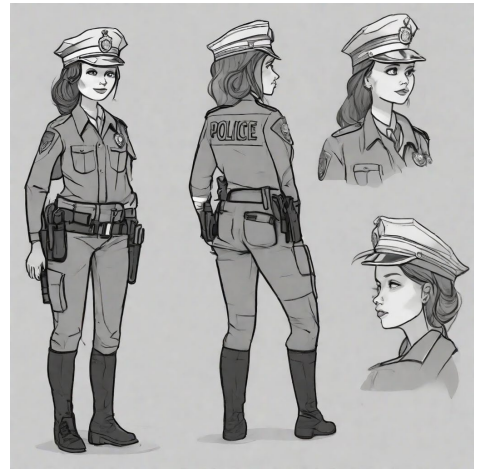


“ Gardner...”

“ Scientist ...”

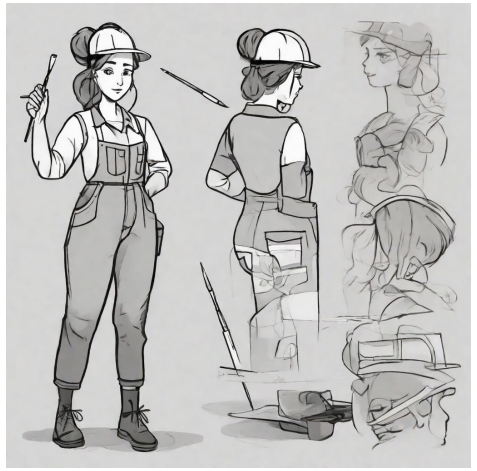


“ Police woman...”

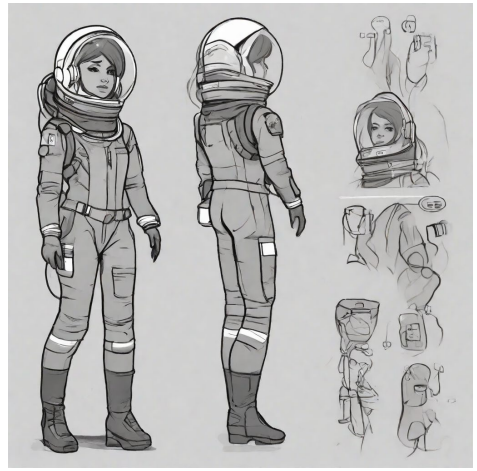


“ Saxophone player...”

“Painter...”



“Astronaut...”



“Taxi Driver...”

“ ...sketch, character sheet.”

Style Aligned generation of Synthetic Images

“Firewoman...”

“ Gardner...”

“ Scientist ...”

“ Police woman...”

“ Saxophone player...”

“Painter...”

“Astronaut...”

“Taxi Driver...”

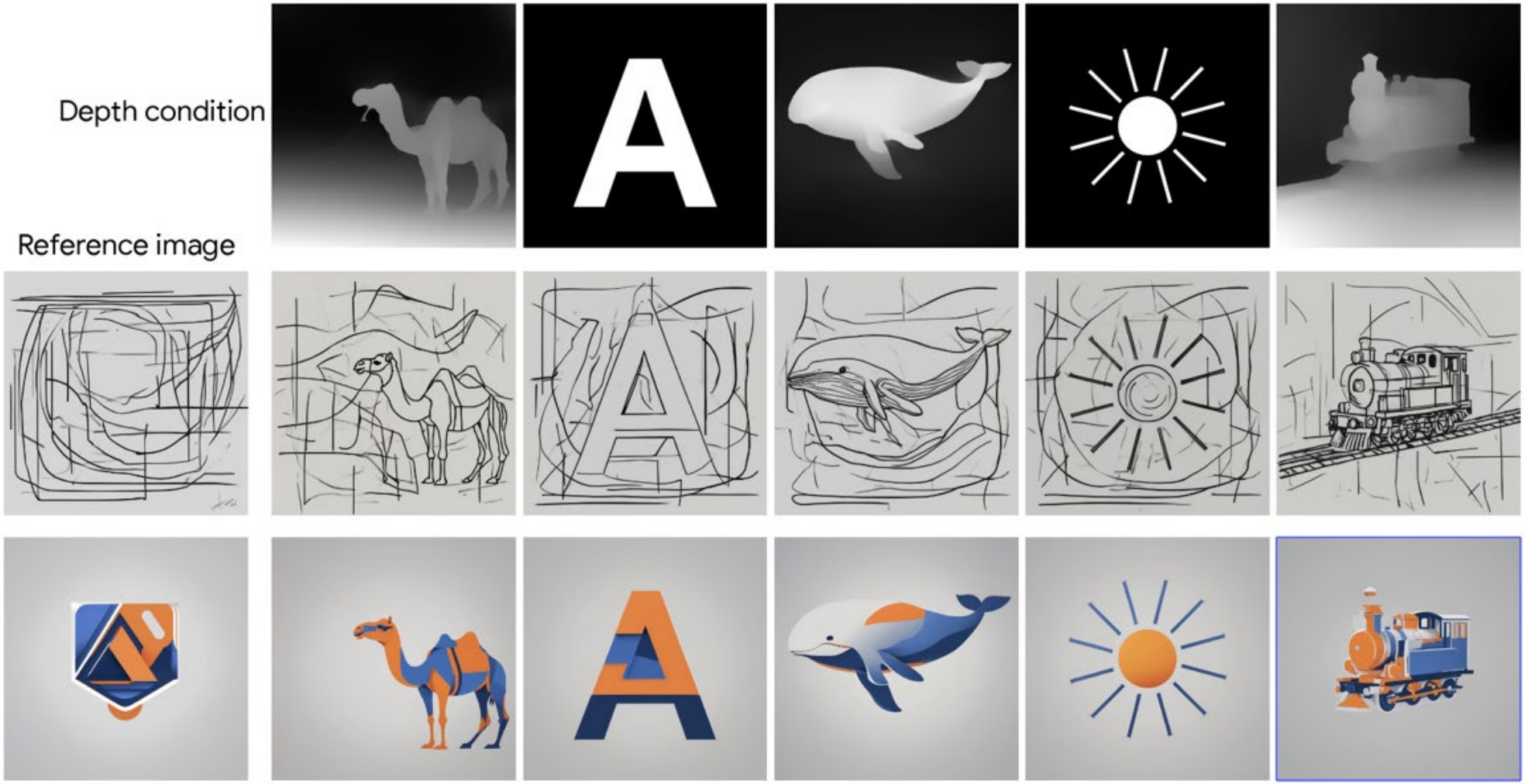
“ ...in minimal flat design illustartion.”

Style Aligned generation from an Input Image

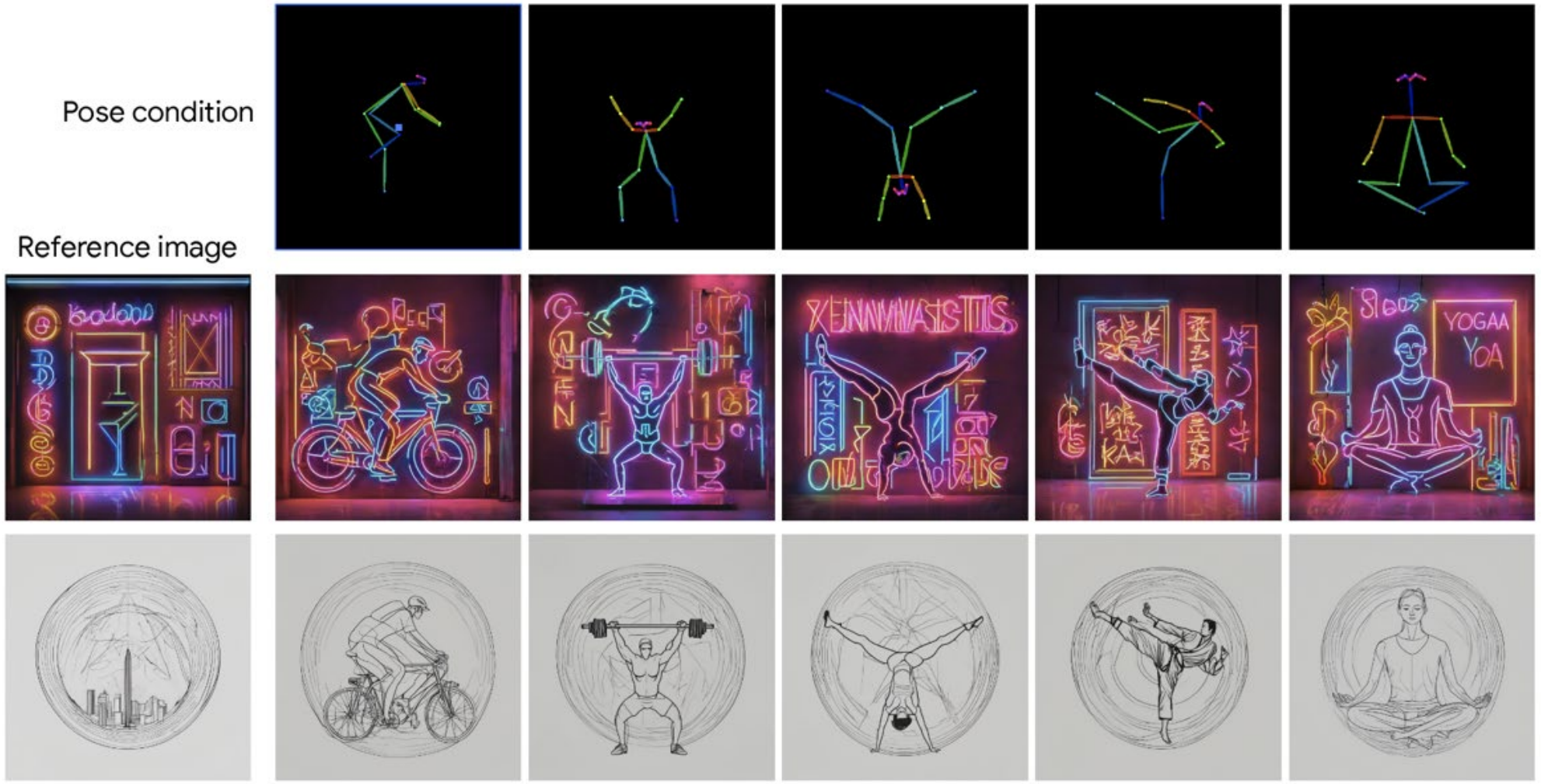
Style Aligned generation from an Input Image

StyleAligned with other methods

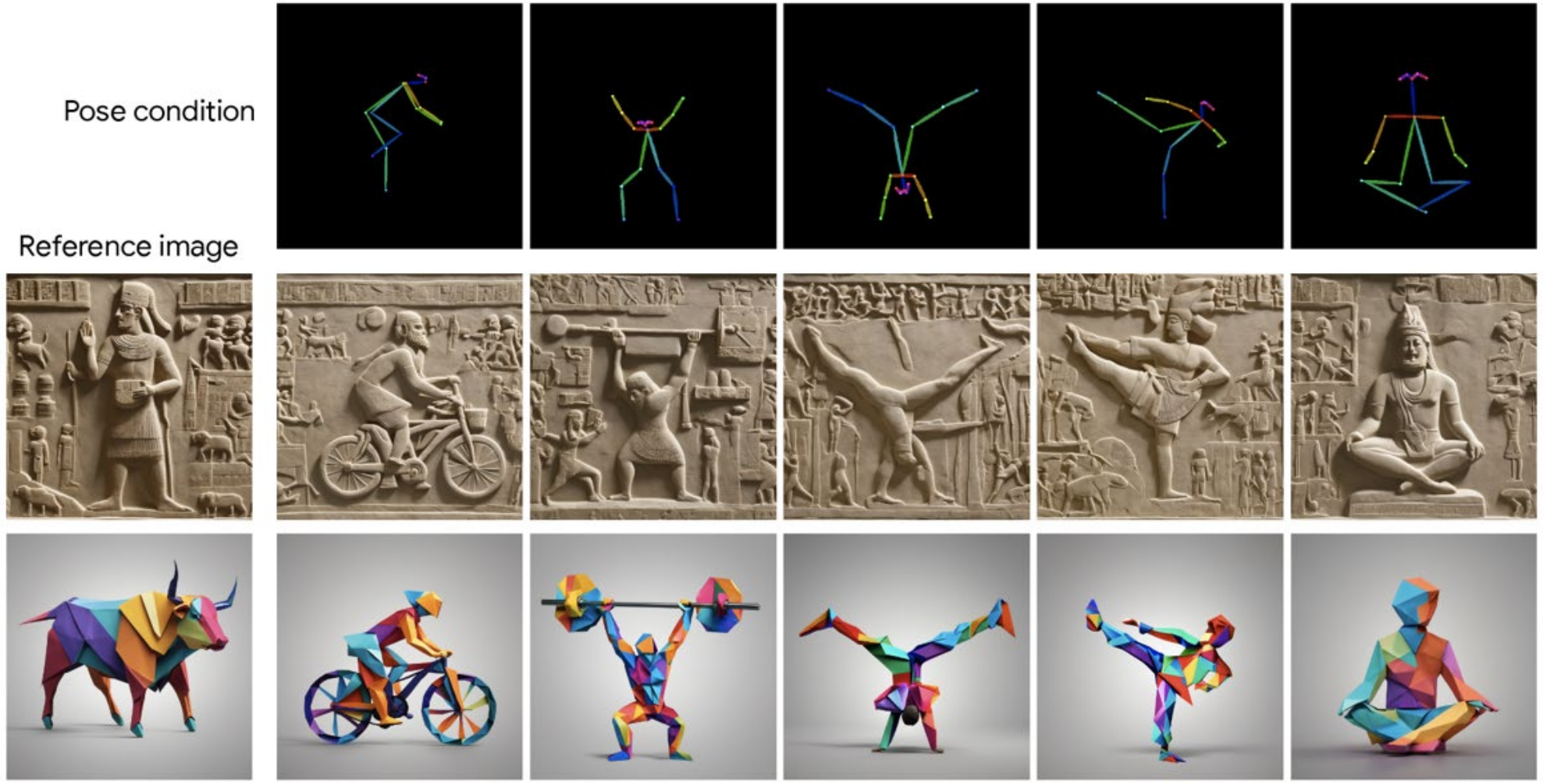
ControlNet + StyleAligned



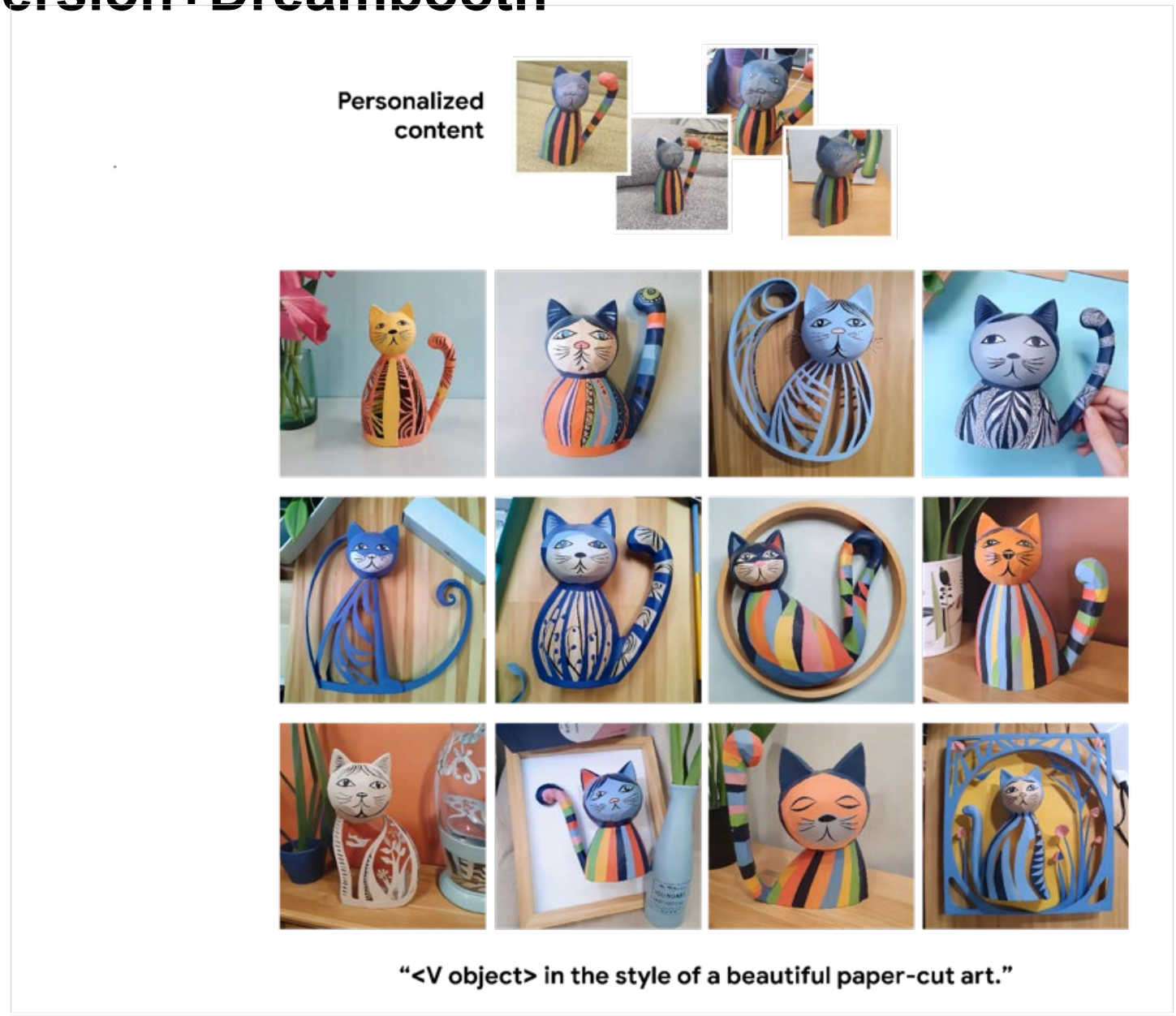
ControlNet + StyleAligned



ControlNet + StyleAligned



Textual Inversion+Dreambooth

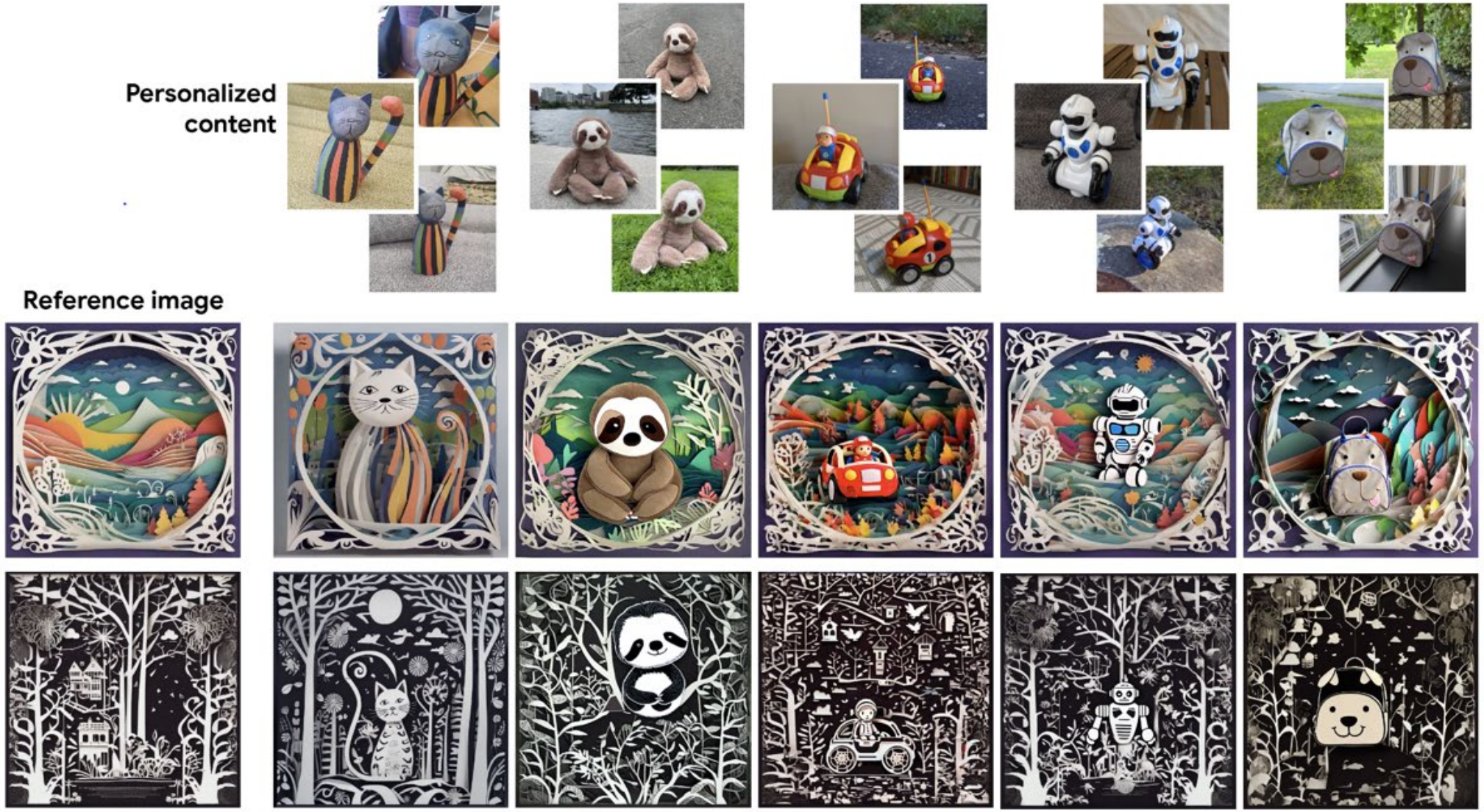


+ StyleAligned

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

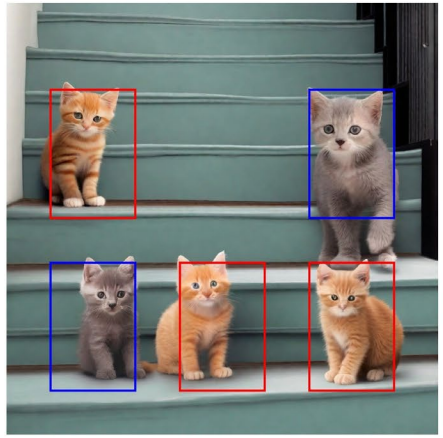
Right reference

Left Reference

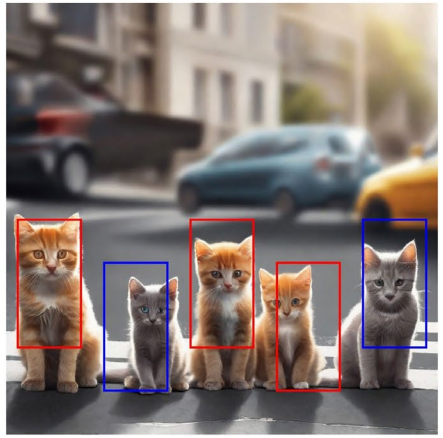
Right Reference

Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation

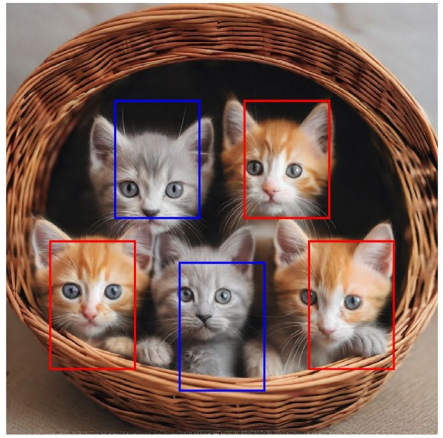
"3 ginger kittens and 2 gray kittens..."



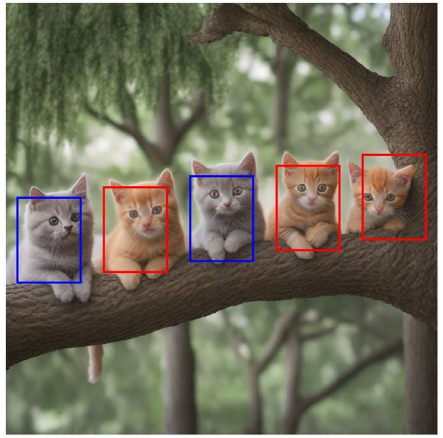
"... on the stairs"



"... on the street"



"... in a basket"



"... on a tree"

Omer Dahary, Or Patashnik, Kfir Aberman, Daniel Cohen-Or

Misalignment in Text-to-Image Generation

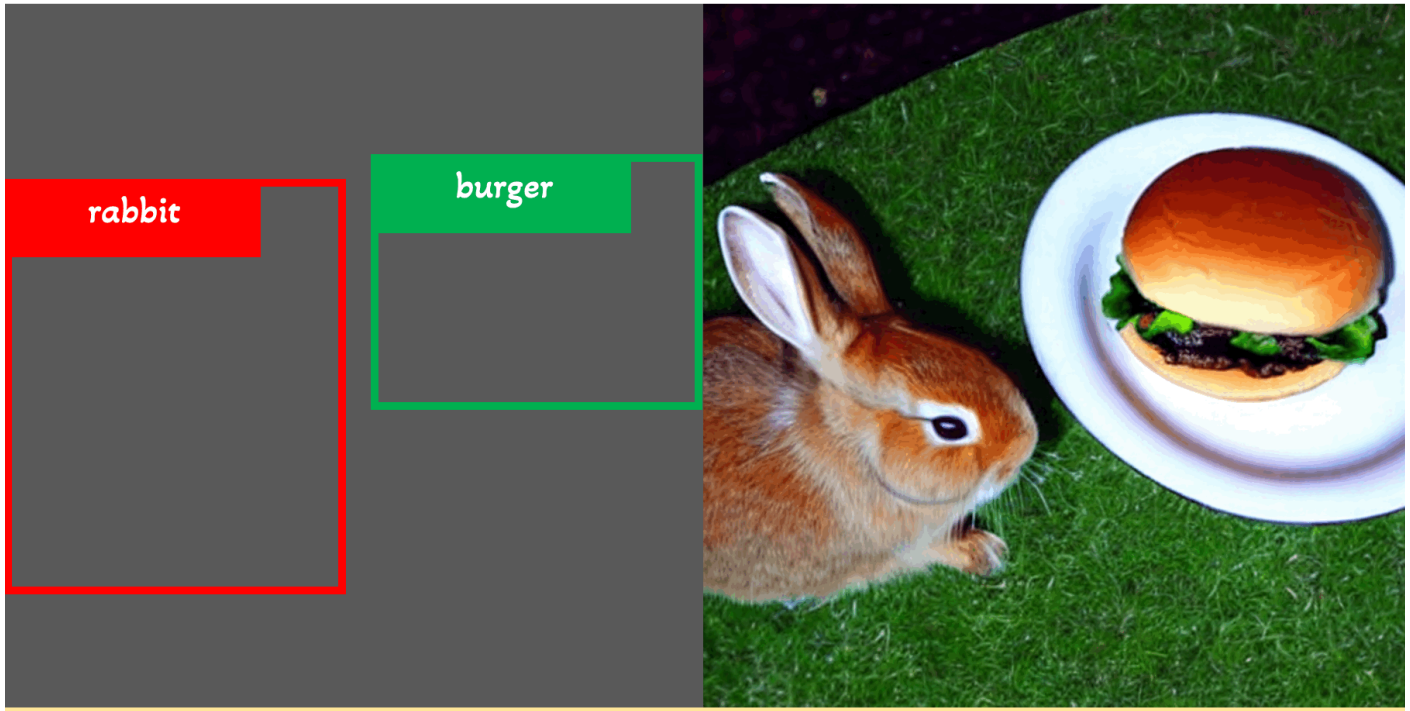
"3D Pixar animation of a **cute unicorn** and a **pink hedgehog** and a **nerdy owl** traveling in a **magical forest**."

Catastrophic neglect

Subject fusion

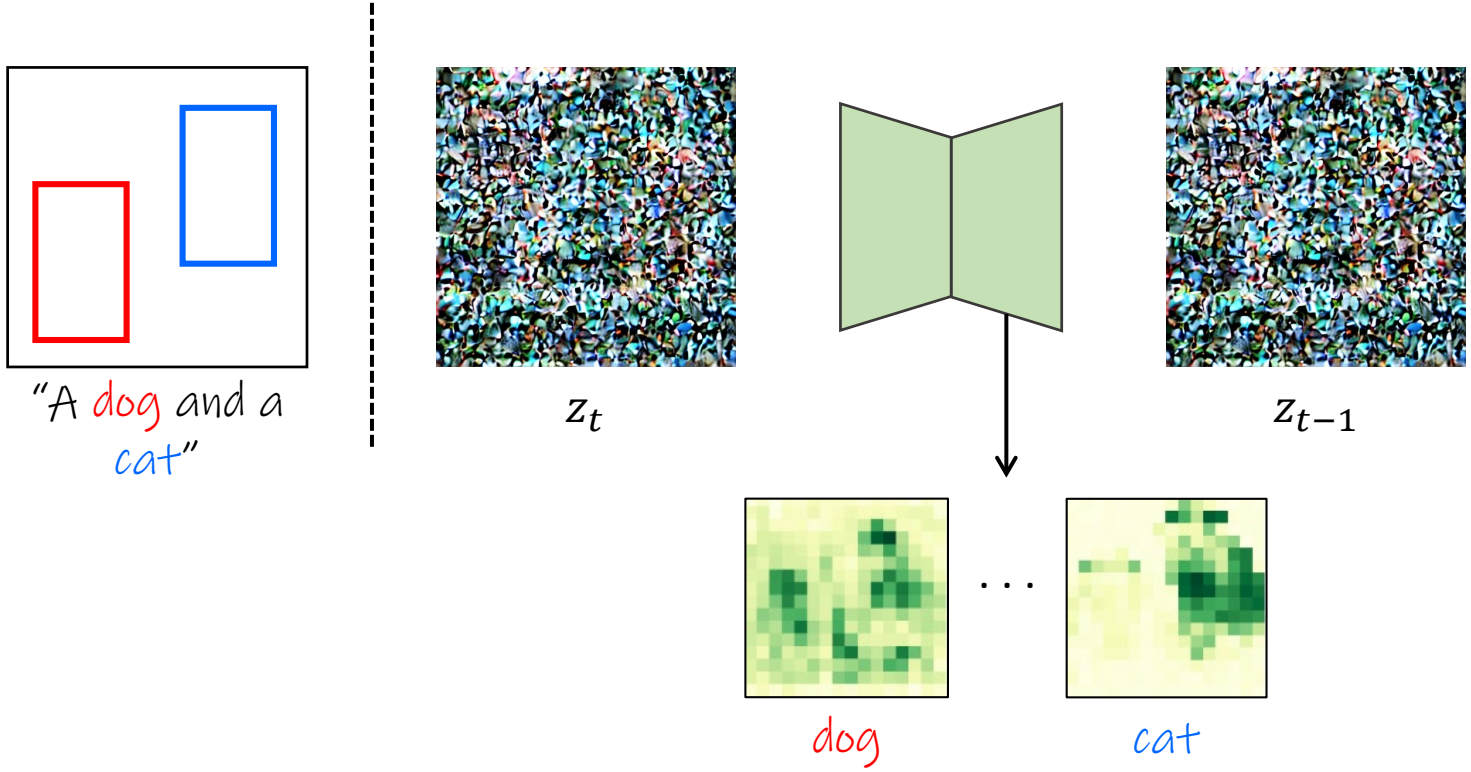
Incorrect attribute binding

Layout-guided Text-to-Image Generation



Caption: A rabbit is about to eat a burger

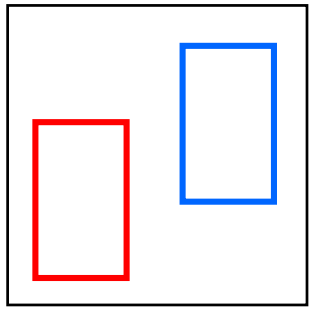
Latent Optimization For Layout Guidance

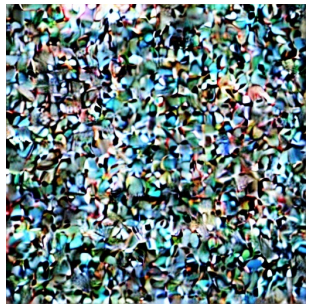


Attend and Excite [Chefer et al., SIGGRAPH 2023]

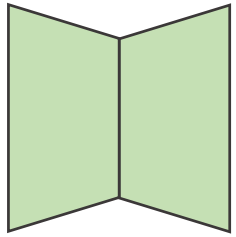
Layout Guidance [Chen et al., WACV 2023]

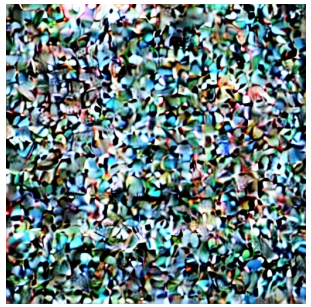
Latent Optimization For Layout Guidance





z_t

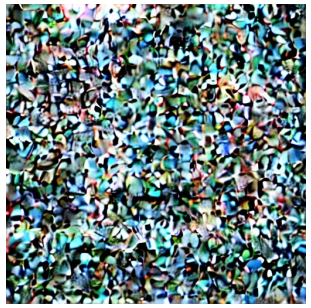




z_{t-1}

$$z_t' = z_t - \alpha \nabla_{z_t} \mathcal{L}$$

$\mathcal{L}$ = Attention of each subject is in its box

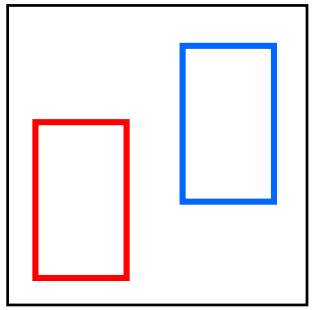


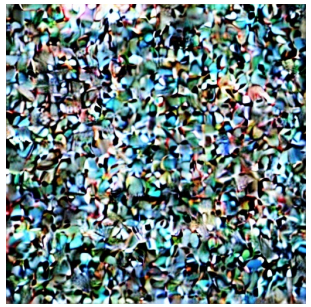
z_t'

Attend and Excite [Chefer et al., SIGGRAPH 2023]

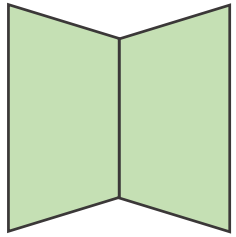
Layout Guidance [Chen et al., WACV 2023]

Latent Optimization For Layout Guidance



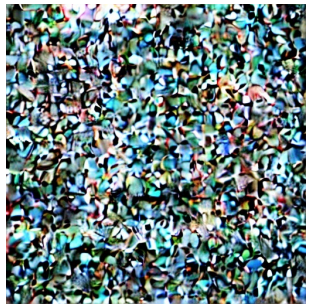


z_t

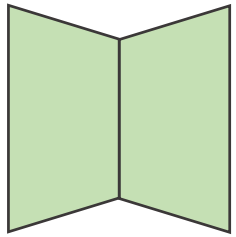


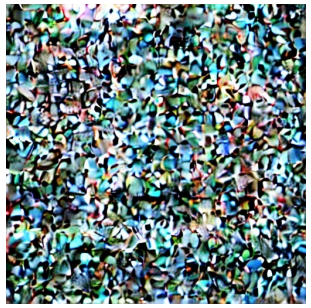
$z_t' = z_t - \alpha \nabla_{z_t} \mathcal{L}$

$\mathcal{L}$ = Attention of each subject is in its box



z_t'





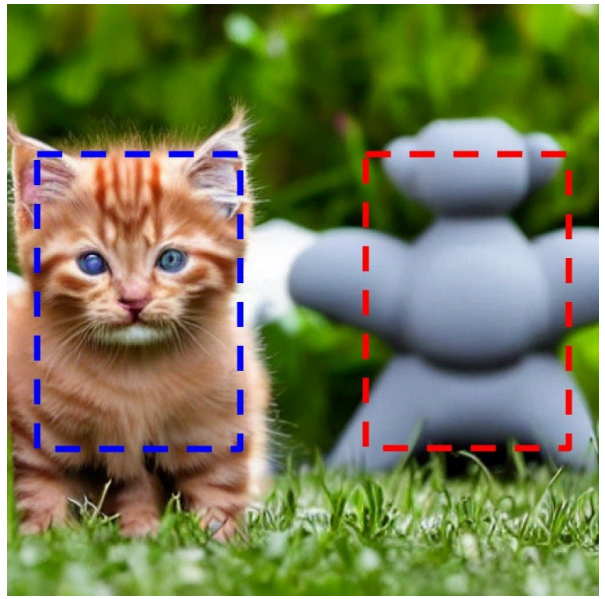
z_{t-1}

Attend and Excite [Chefer et al., SIGGRAPH 2023]

Layout Guidance [Chen et al., WACV 2023]

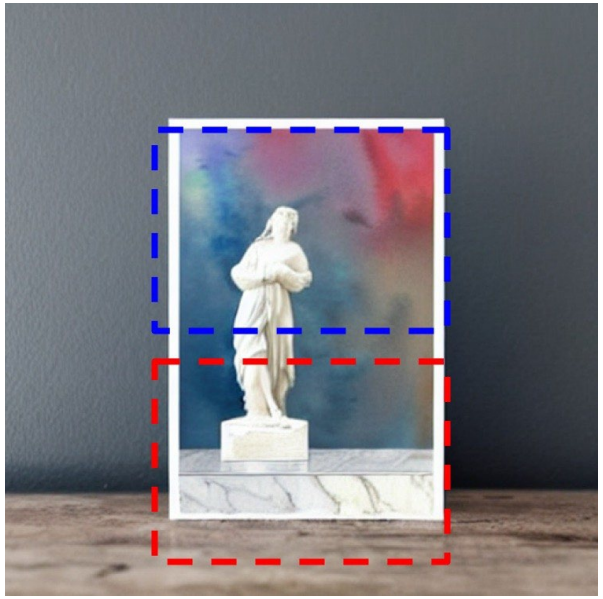
Misalignment in Layout Conditioned Text-to-Image Generation

"A ginger kitten and a gray puppy"



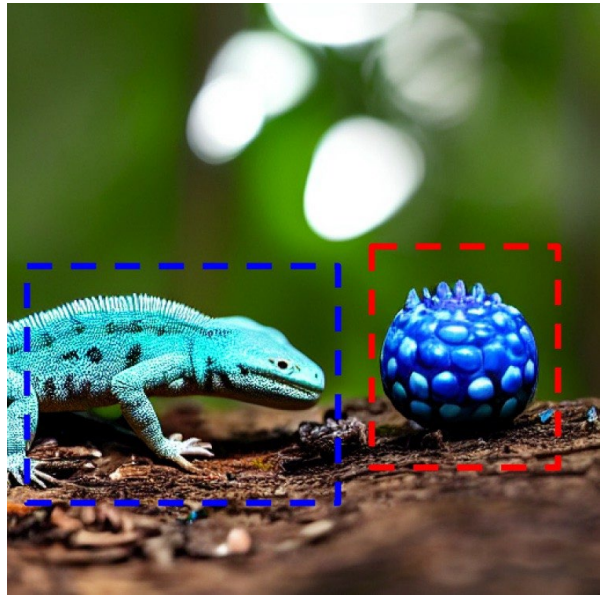
Catastrophic neglect

"A watercolor painting and a marble statue"



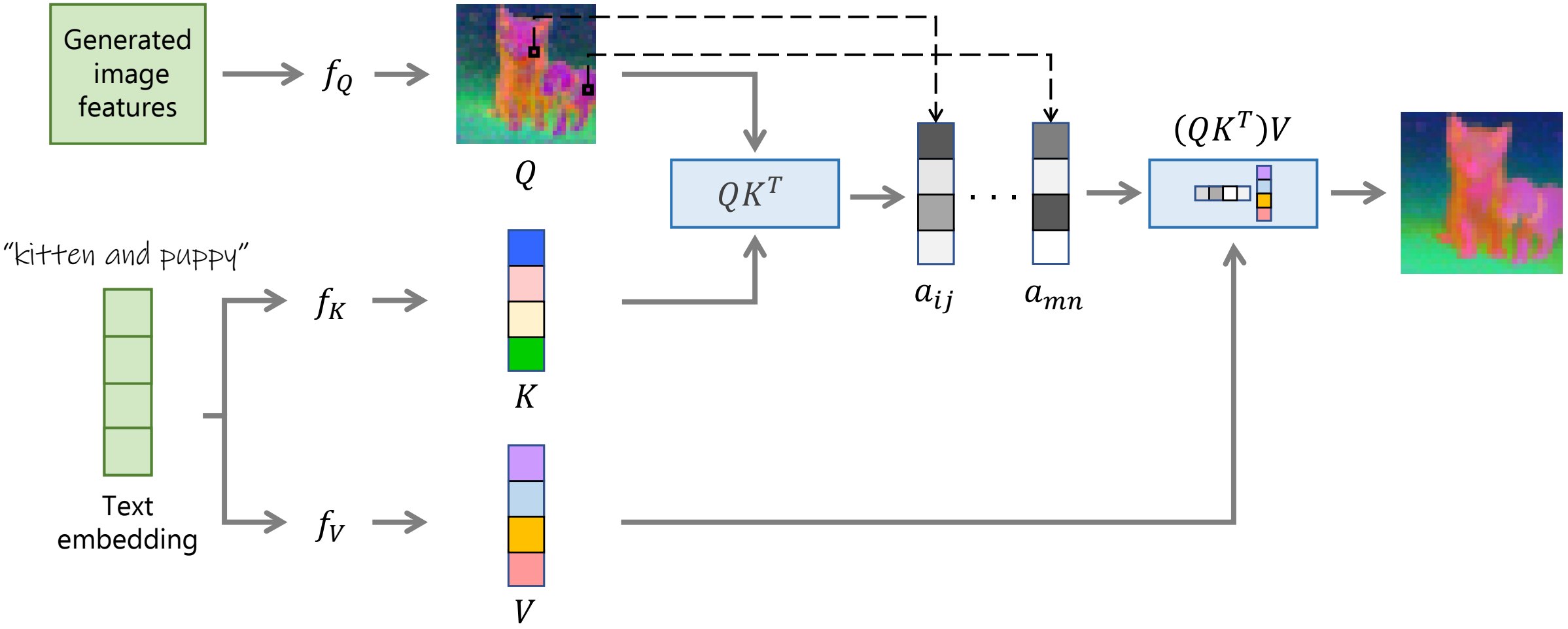
Subject fusion

"A spotted lizard and a blue fruit"

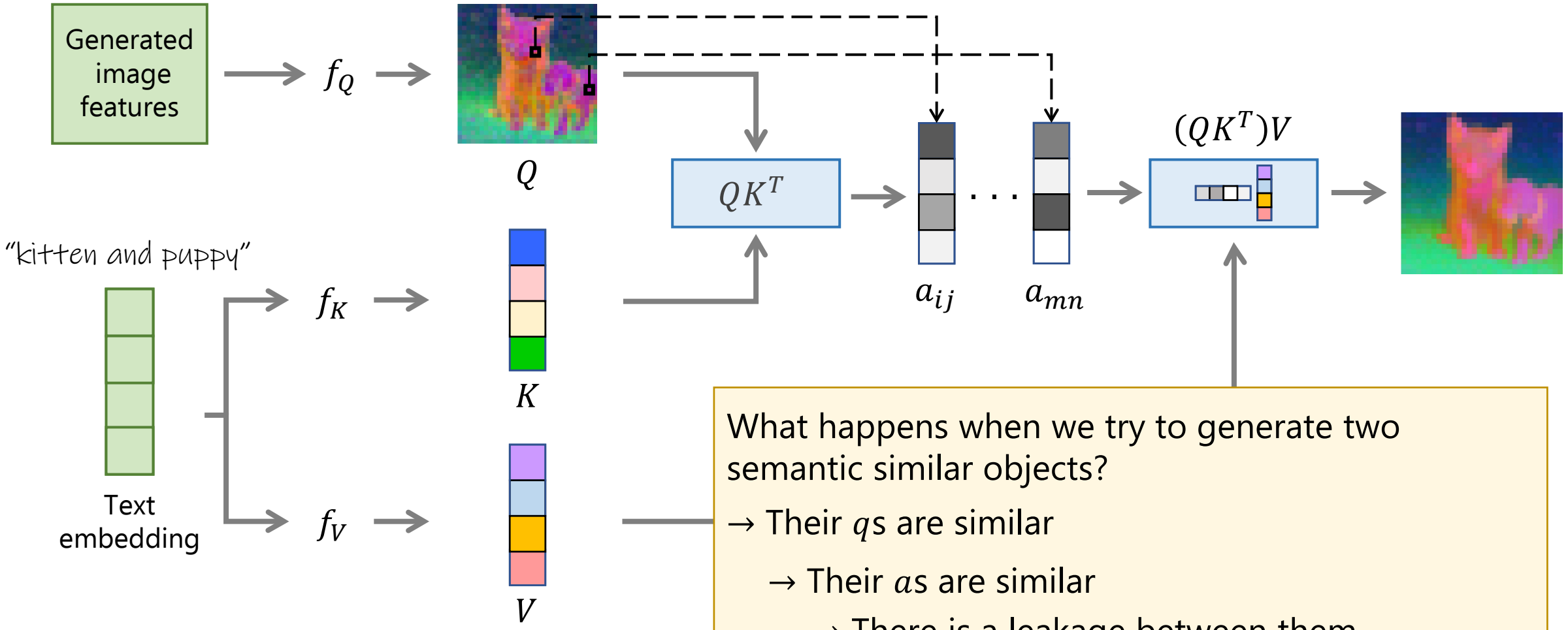


Incorrect attribute binding

Cross-Attention Layers



Cross-Attention Layers

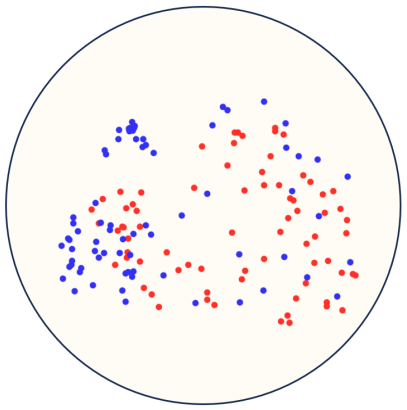


Leakage In Cross-Attention Layers

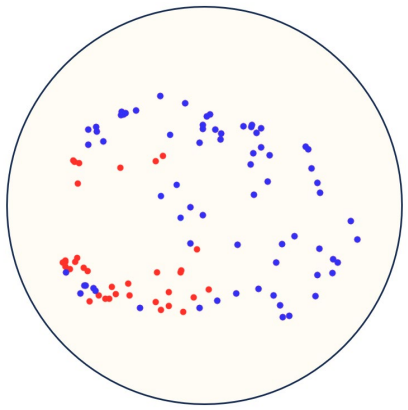
"A hamster"

"A squirrel"

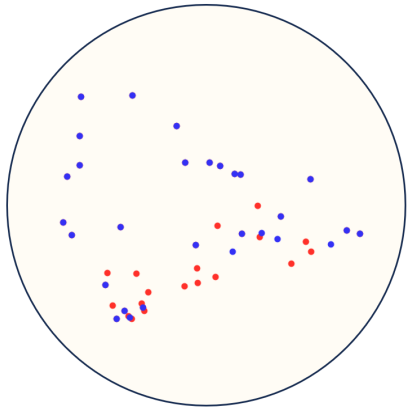
← "A hamster and a squirrel" →



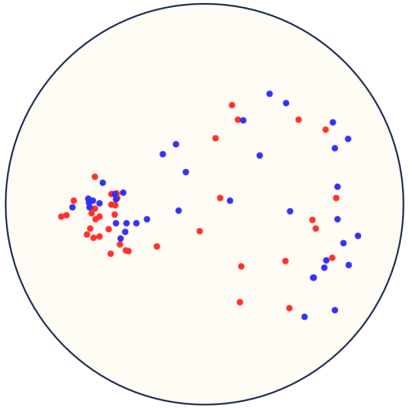
Stable Diffusion



Stable Diffusion



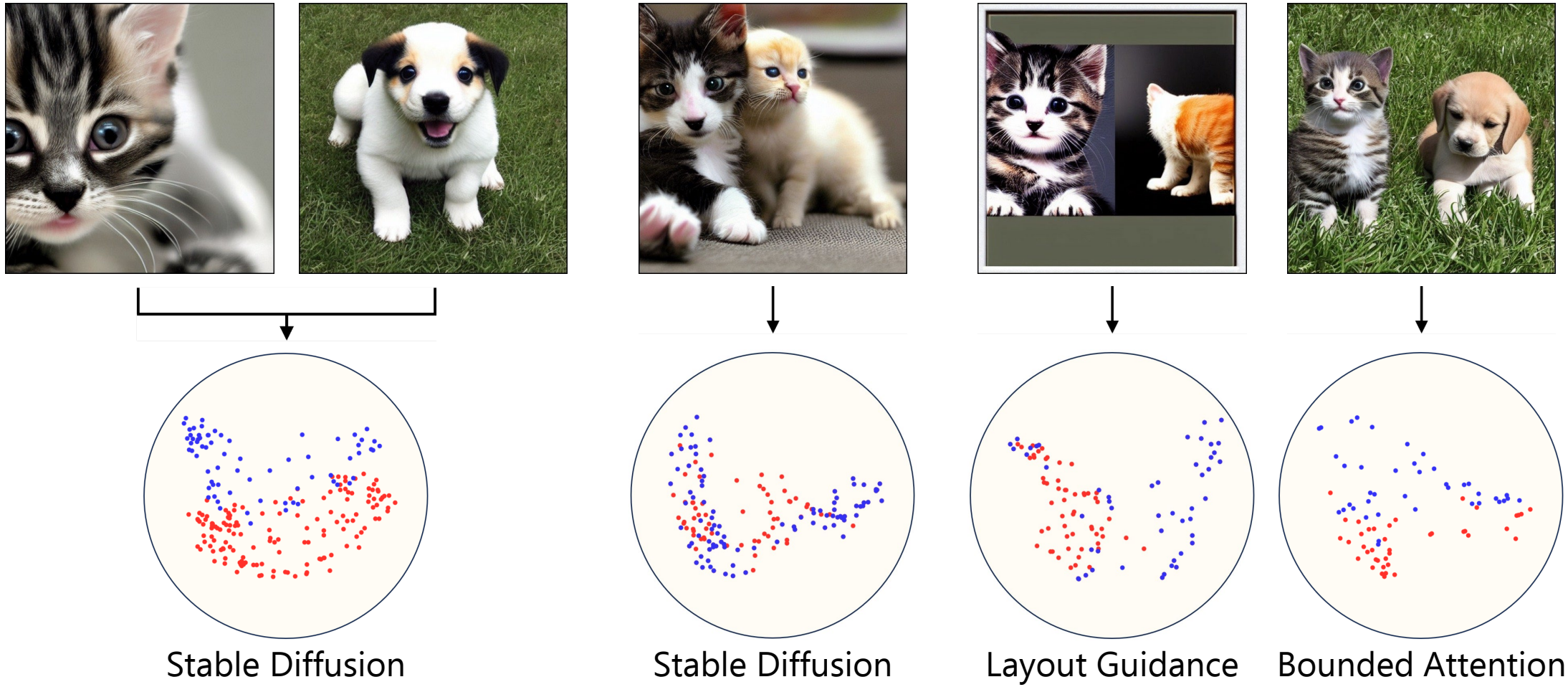
Layout Guidance



Bounded Attention

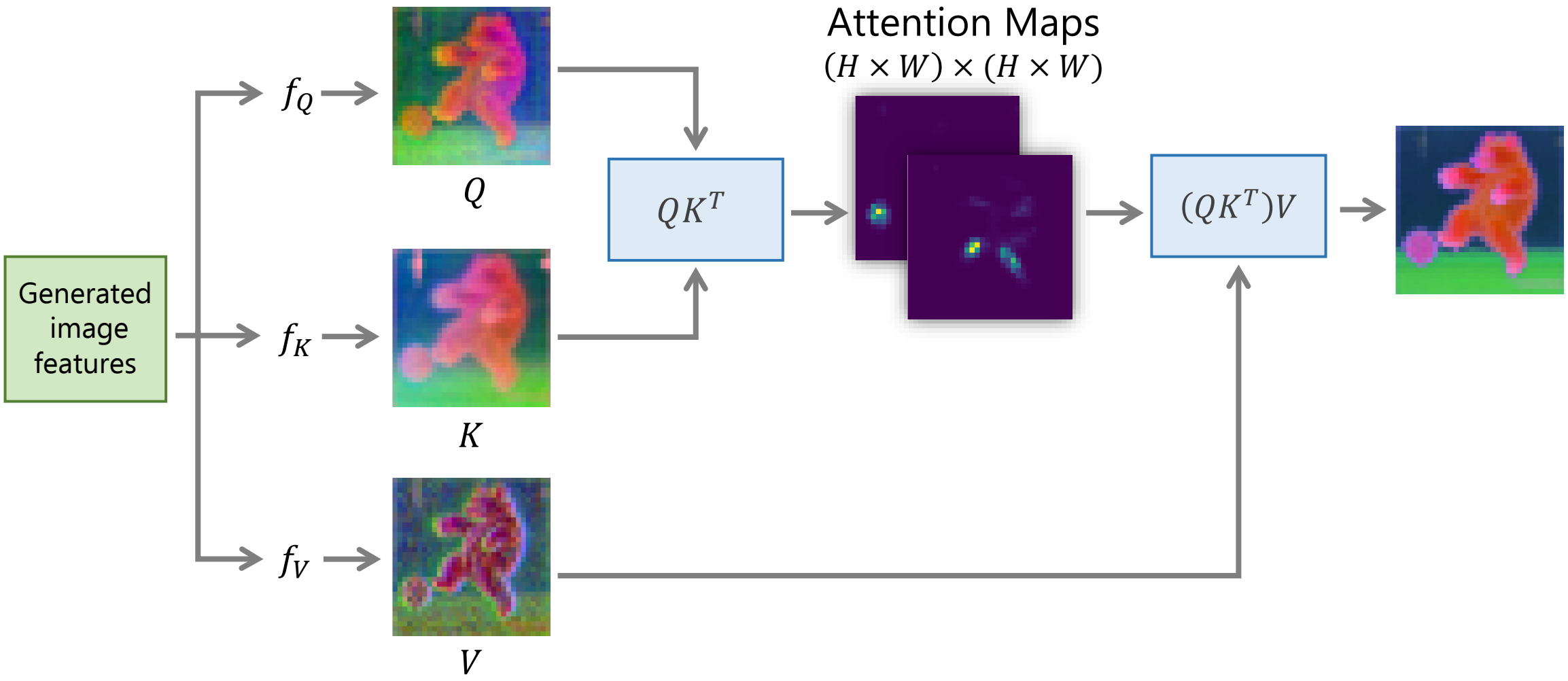
Leakage In Cross-Attention Layers

"A kitten" "A puppy" ← "A kitten and a puppy" →

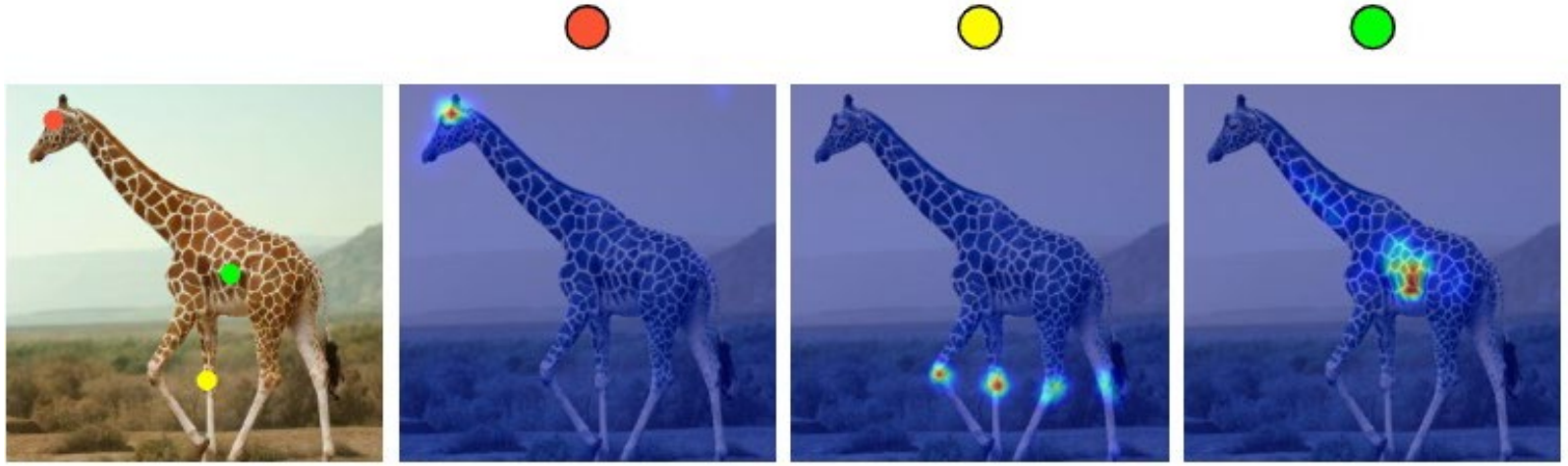


Stable Diffusion Stable Diffusion Layout Guidance Bounded Attention

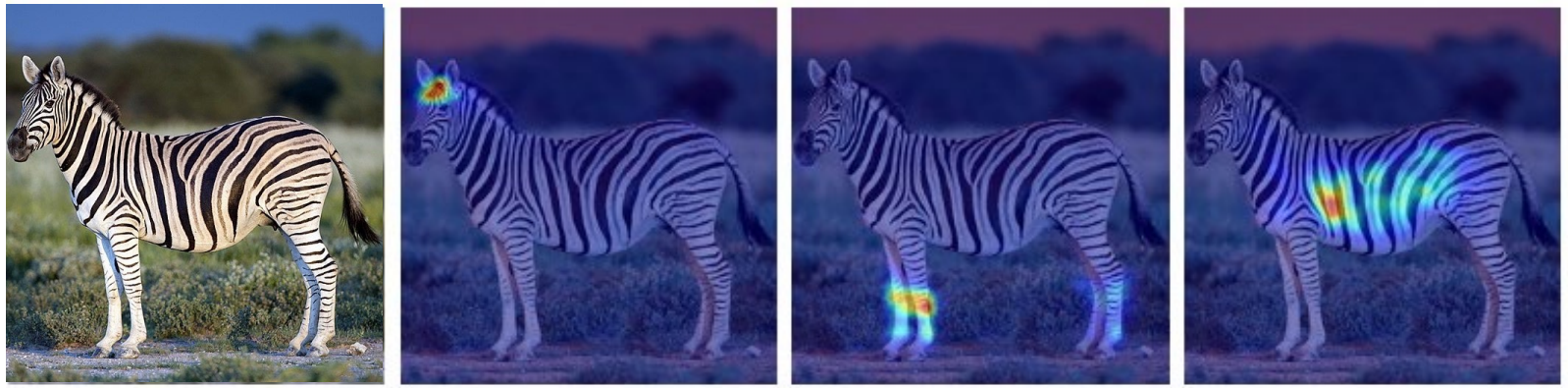
Self-Attention Layers



The Roles of the Queries, Keys, and Values



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

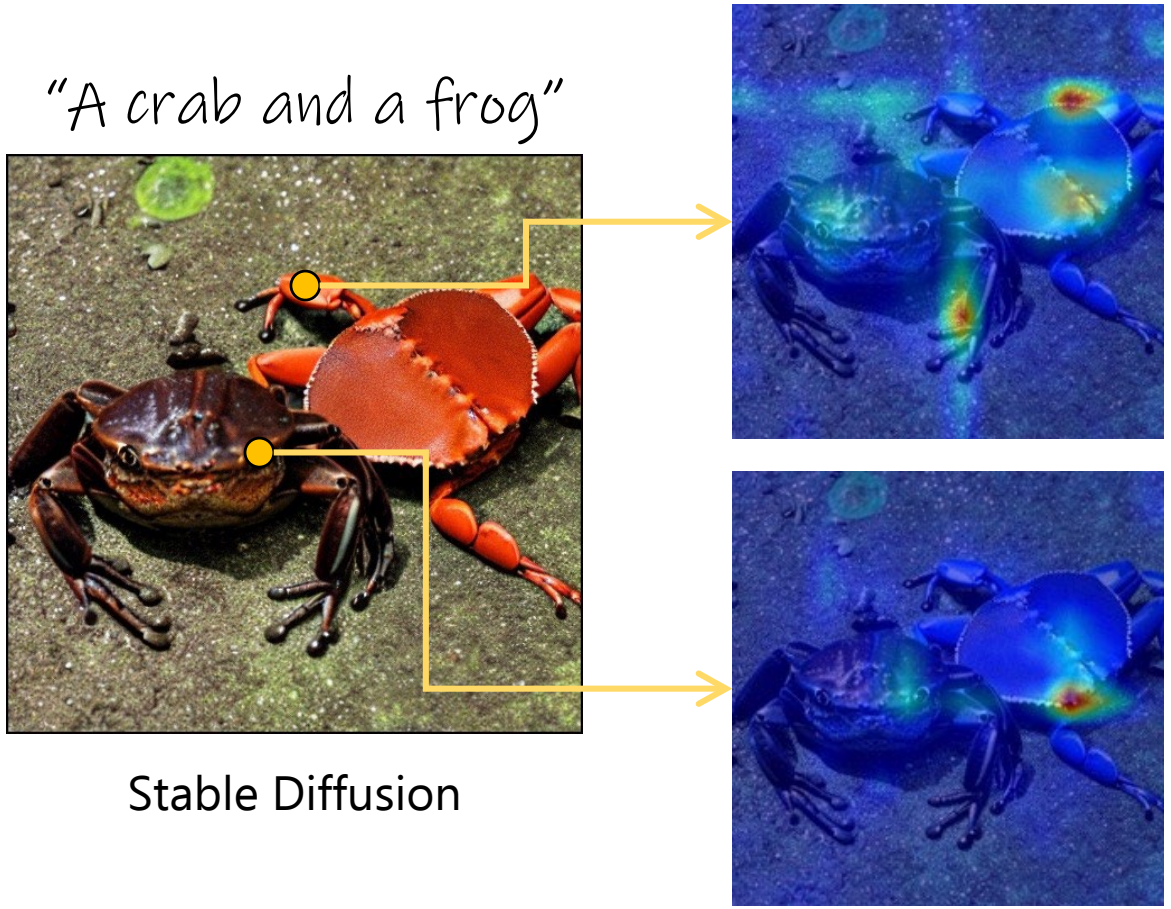


$$Q_{struct} \cdot K_{app}^T$$

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

Leakage In Self-Attention Layers

"A crab and a frog"



Stable Diffusion