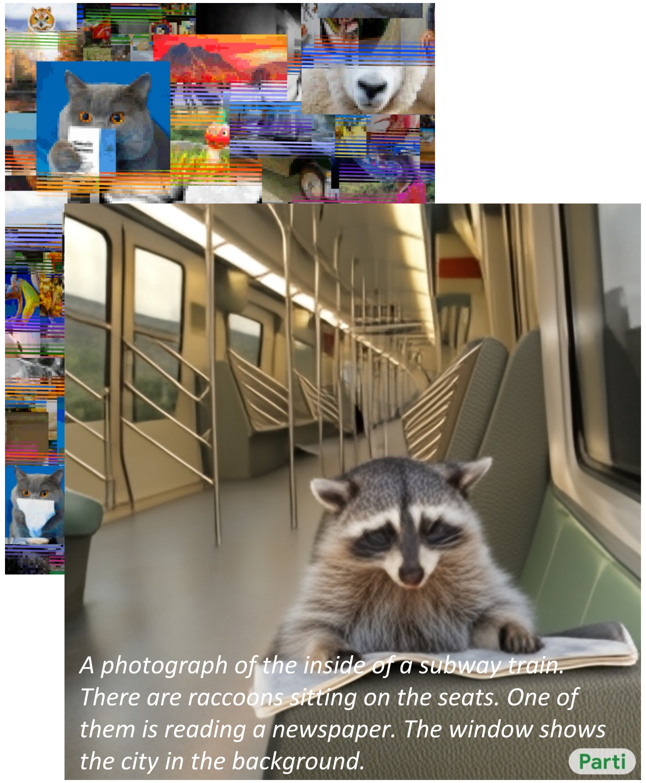
Image Editing with Optimization (part II)

Jun-Yan Zhu

16-726 Spring 2025

Large-scale Text-to-Image Models

Diffusion models
(DALL-E 2, Imagen, SD)



Autoregressive models
(Image GPT, Parti)

GANs, Masked GIT
(GigaGAN, MUSE)

Limitations of Text-to-Image Models

Linguistic bottleneck: not everything can be described by text

Data bottleneck: many things are not included in the dataset:

1. Not in the public domains (e.g., personal concepts)
2. Have not been created (e.g., new concepts)

Text-to-image isn't perfect…

Stable Diffusion

Photo of a moongate

Text-to-image isn't perfect…

Stable Diffusion

Actual moongate images

`Photo of a moongate`

Text-to-image isn't perfect…

Actual moongate images

Stable Diffusion

Photo of a moongate

Customization

Actual moongate images

Stable Diffusion

Photo of a moongate

Customization

Actual moongate images

Customized Diffusion

Photo of a moongate

Unseen contexts

Actual moongate images

Customized Diffusion

Moongate in the middle of highway

Unseen contexts

Customized Diffusion

Actual moongate images

`Moongate in snowy ice`

Unseen contexts

Actual moongate images

Customized Diffusion

A puppy in front of Moongate

No knowledge of personal concepts

My dog, Stark

Stable Diffusion

A dark grey color weimaraner dog

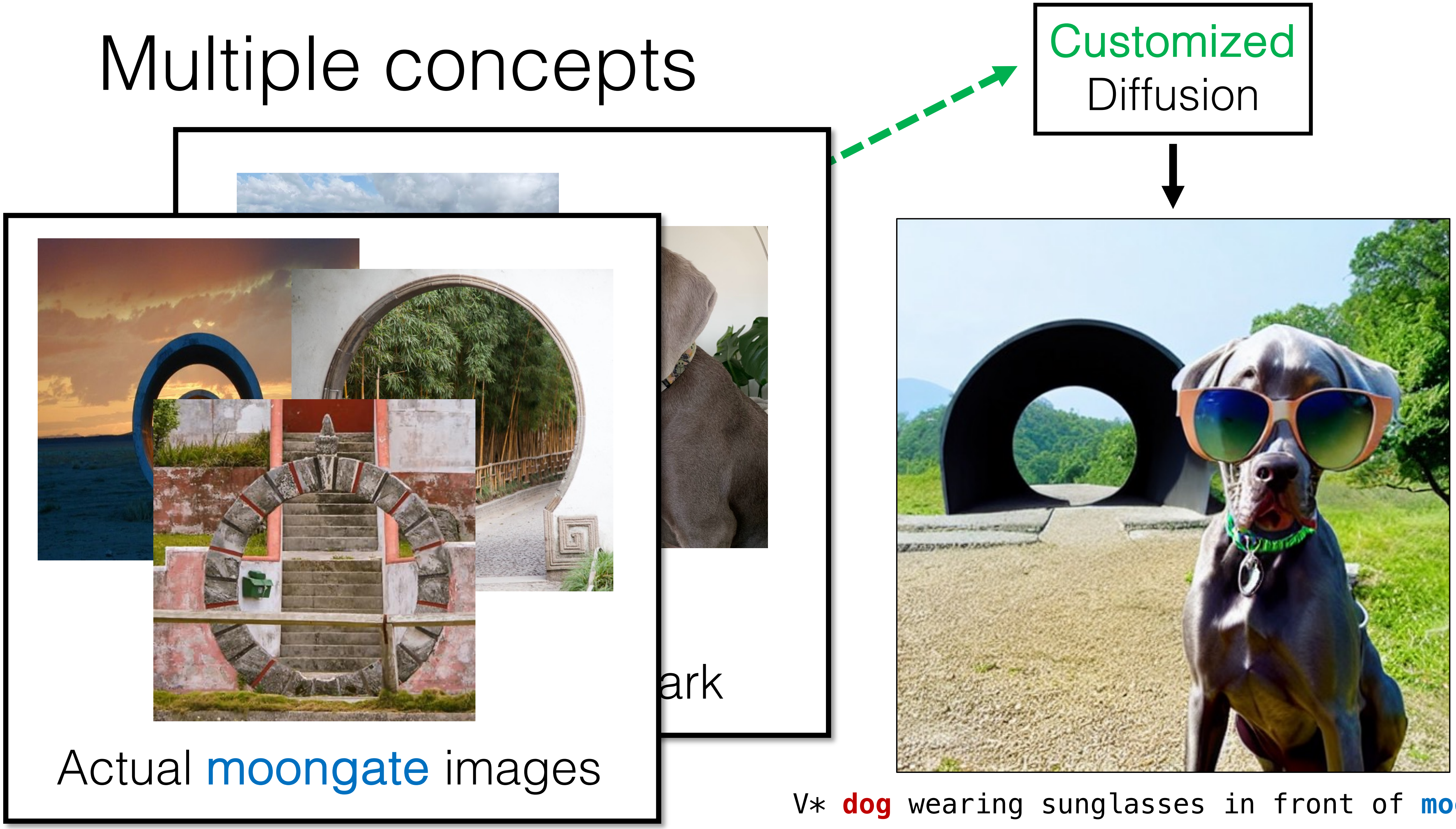
Customization

Jun-Yan's dog, Stark

Customized Diffusion

V* dog wearing sunglasses

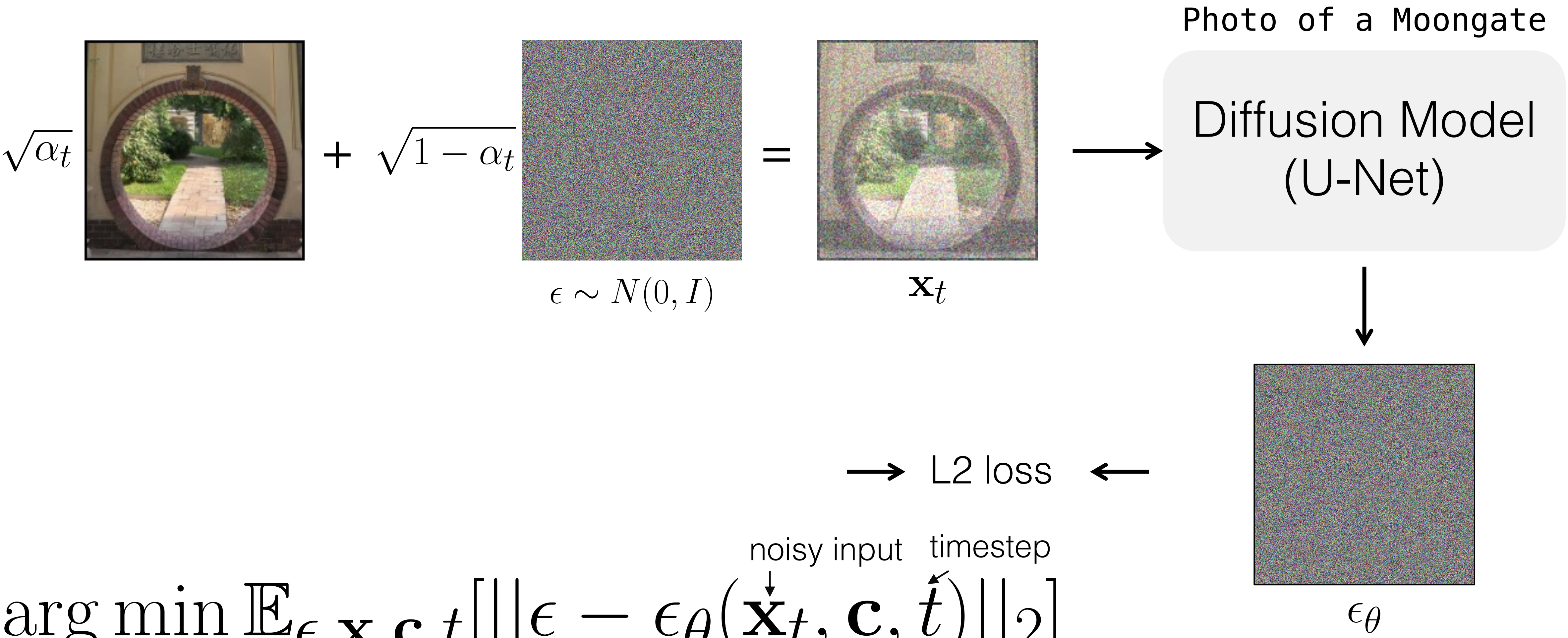
Multiple concepts



V* dog wearing sunglasses in front of moongate

Diffusion Model Quick Recap

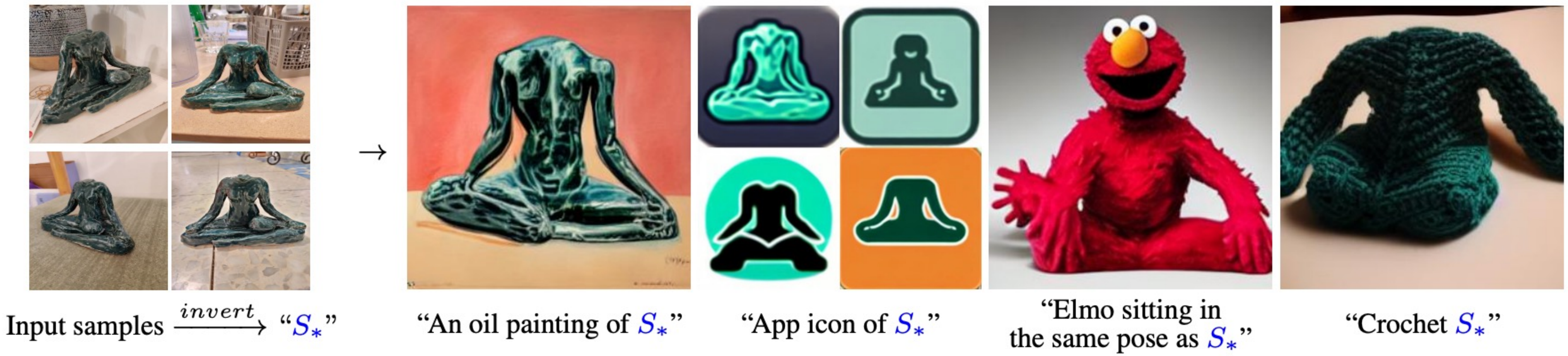
Diffusion model training



$$\arg\min_{\theta} \mathbb{E}_{\epsilon,\mathbf{x},\mathbf{c},t}[||\epsilon - \underset{\text{U-Net}}{\epsilon_\theta}(\overset{\text{noisy input}}{\mathbf{x}_t}, \underset{\text{caption}}{\mathbf{c}}, \overset{\text{timestep}}{t})||_2]$$

Which parts shall we customize?

Textual Inversion: Optimizing Text Embedding

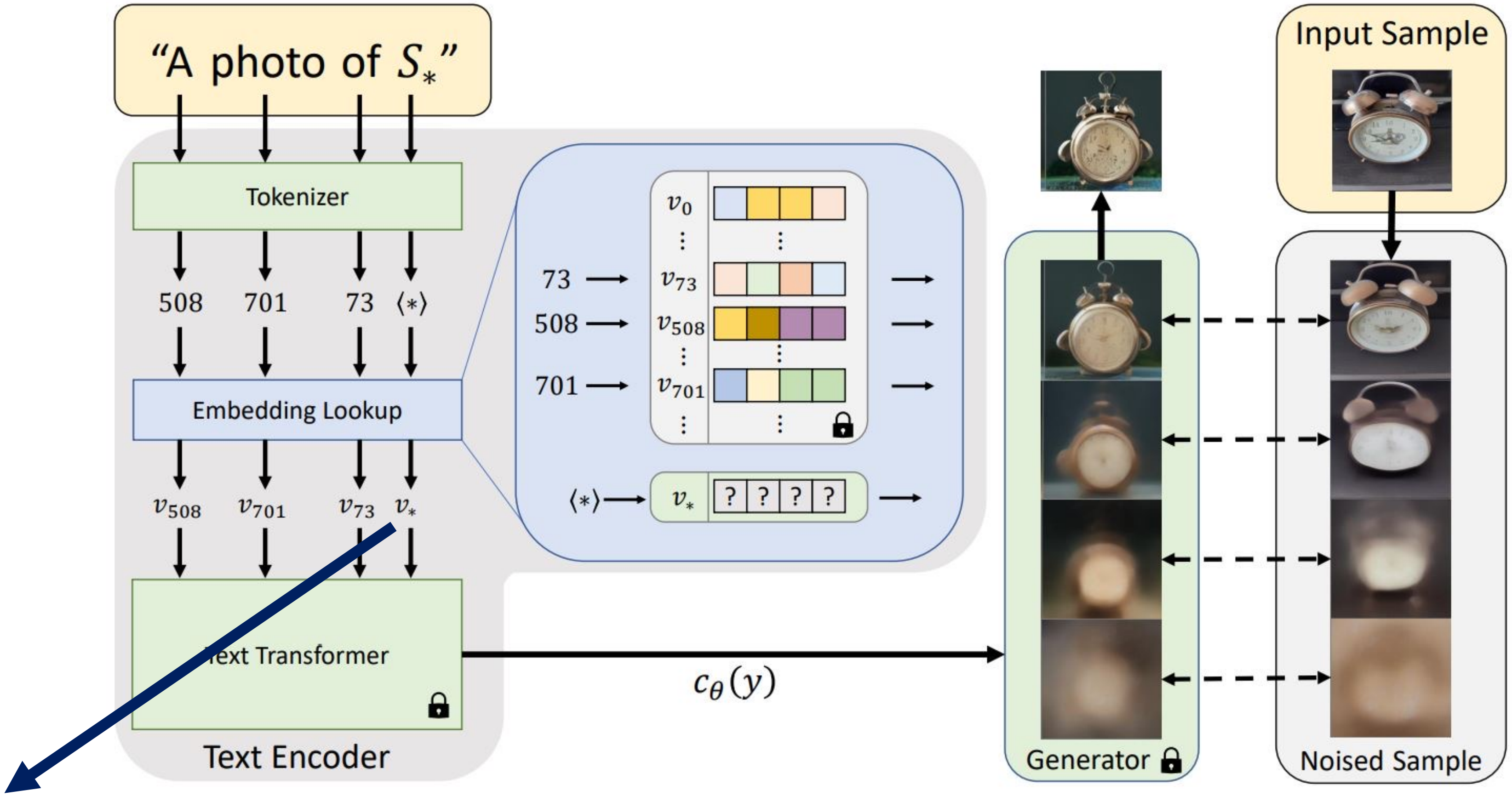


[Rinon Gal et al., ICLR 2023]

Textual Inversion: Optimizing Text Embedding

[Rinon Gal et al., ICLR 2023]

Textual Inversion: Optimizing Text Embedding



$$v^* = \arg\min_v \mathbb{E}_{\epsilon,\mathbf{x},\mathbf{c},t}[||\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)||_2]$$

GANs inversion [Zhu et al., 2016] and soft prompting [Lester et al., 2021]

[Rinon Gal et al., ICLR 2023]

Textual Inversion Results

[Rinon Gal et al., ICLR 2023]

Textual Inversion Results

[Rinon Gal et al., ICLR 2023]

Works well for artistic styles

Input samples

"The streets of Paris in the style of S_*"

"Adorable corgi in the style of S_*"

"Painting of a black hole in the style of S_*"

"Times square in the style of S_*"

[Rinon Gal et al., ICLR 2023]

Cannot preserve object identity

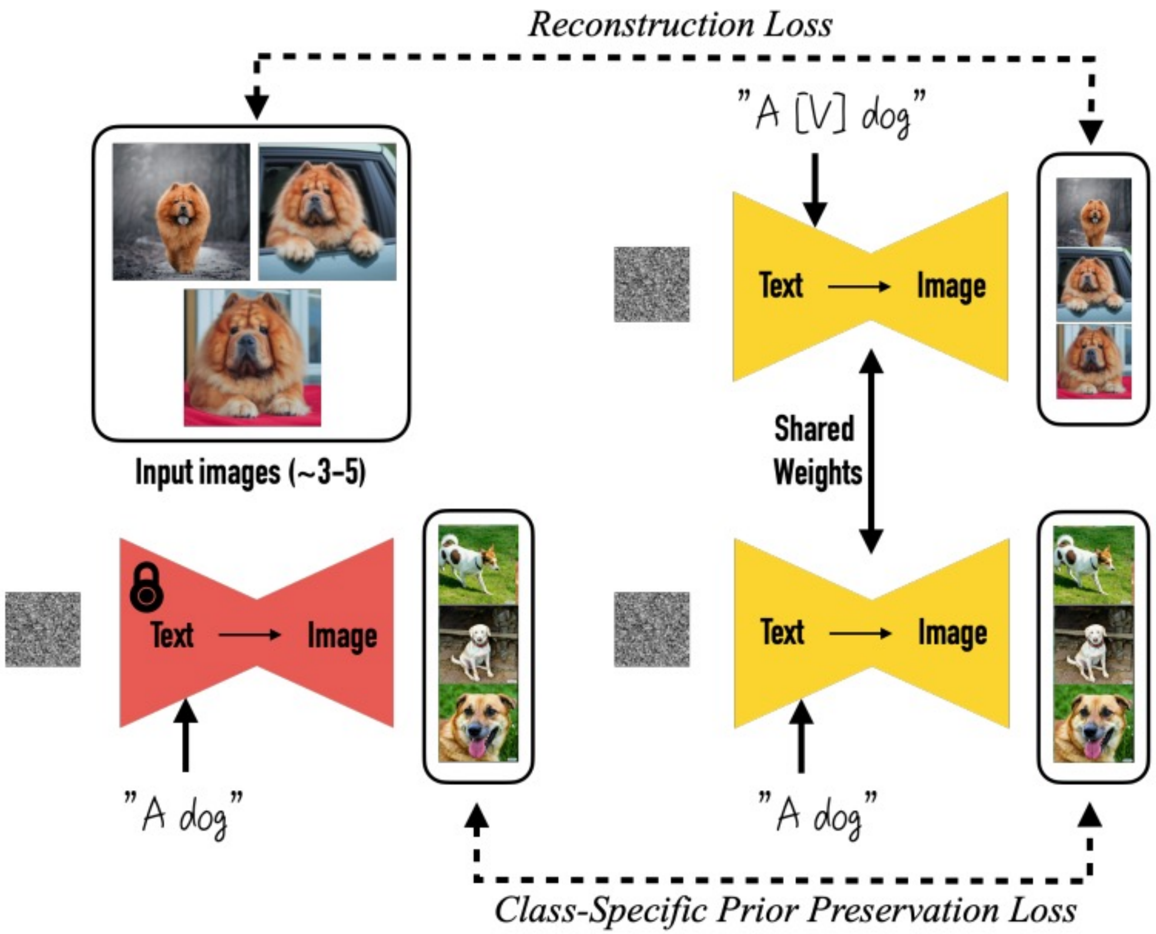
Target images

S* cat swimming in a pool

[Rinon Gal et al., ICLR 2023]

How to improve identity preservation?

DreamBooth: Fine-tuning all the weights



Training Objective

$$\Delta\theta^* = \arg\min_{\Delta\theta} \mathbb{E}_{\epsilon,\mathbf{x},\mathbf{c},t}[||\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)||_2]$$

where $\theta = \theta_0 + \Delta\theta$

Issues (Overfitting)

- Forget to generate subjects of the same class (e.g., dog)
- Reduce output diversity

Regularization

- Add synthetic images of the same class.

Inspired by single-image GAN fine-tuning
GANPaint [Bau et al., 2019], PTI [Roich et al., 2021]

[Nataniel Ruiz et al., CVPR 2023]

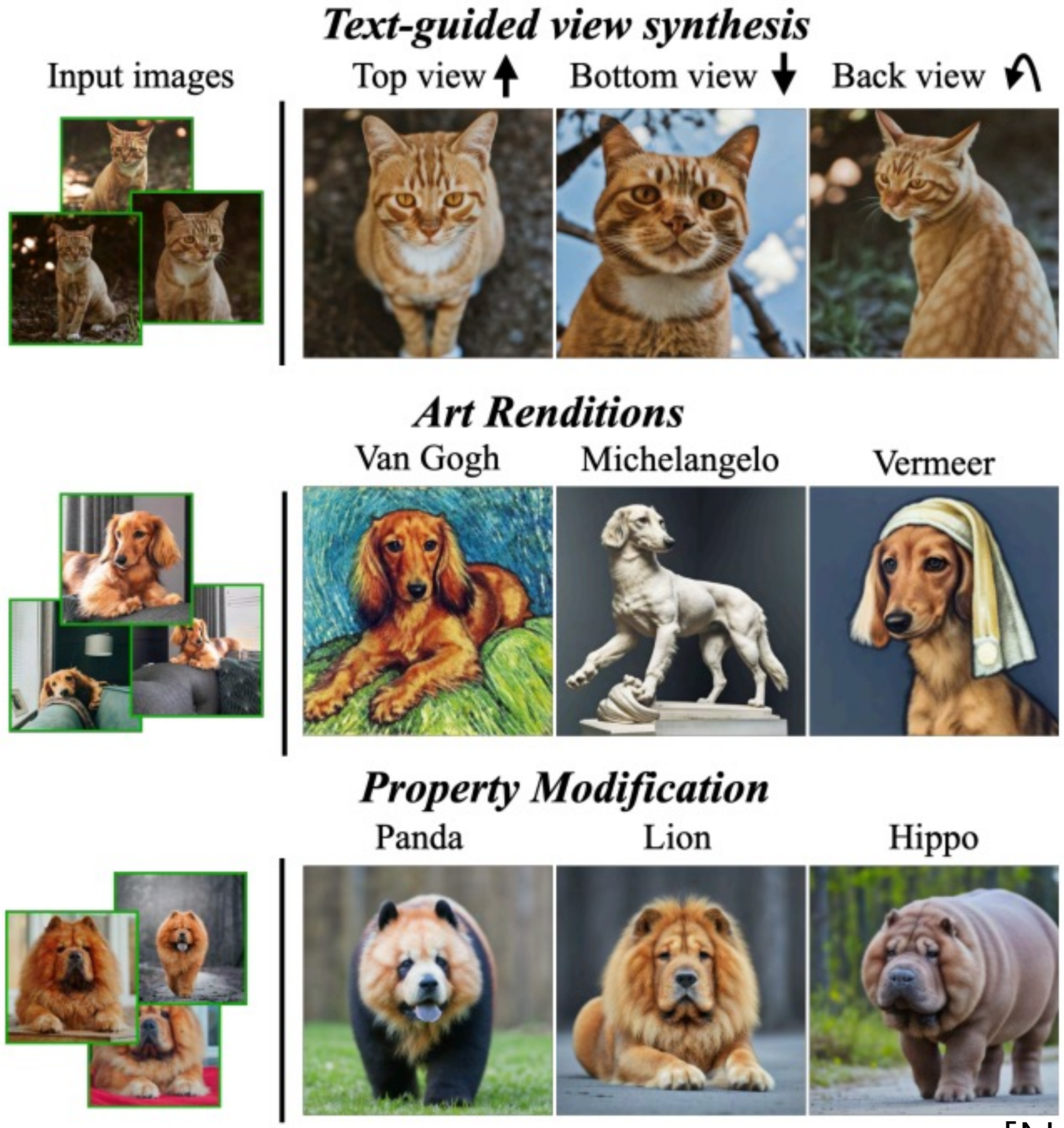
DreamBooth Results

[Nataniel Ruiz et al., CVPR 2023]

DreamBooth Results

[Nataniel Ruiz et al., CVPR 2023]

DreamBooth Applications



[Nataniel Ruiz et al., CVPR 2023]

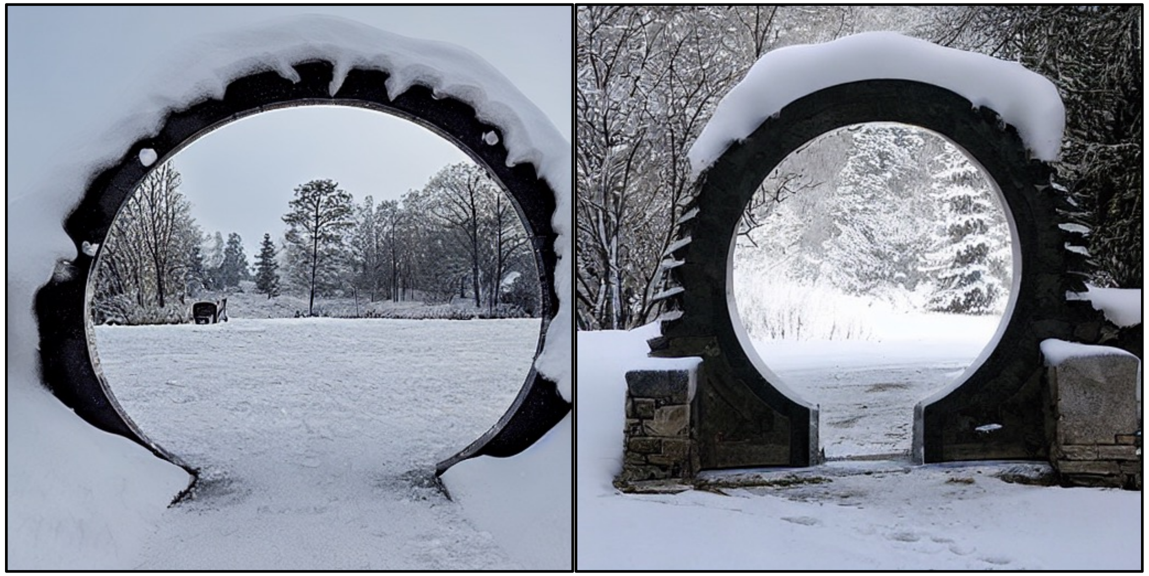
DreamBooth vs. Textual Inversion

[Nataniel Ruiz et al., CVPR 2023]

Fine-tuning all model weights

`Photo of a moongate`

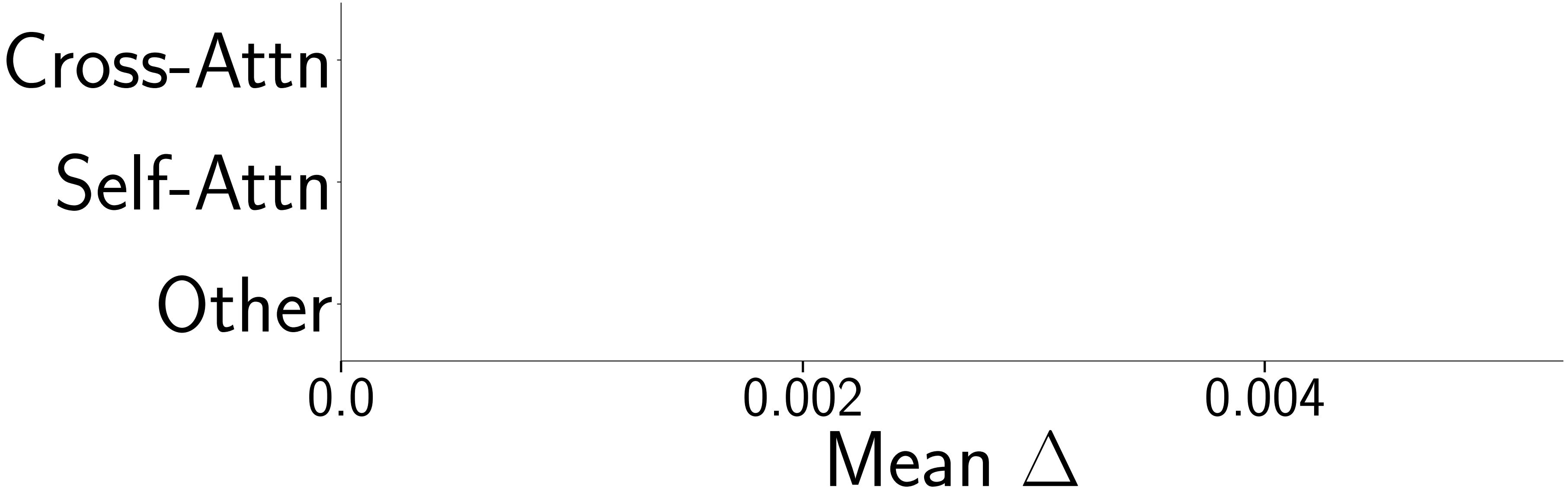
`Moongate in snowy ice`



Storage requirement. 4GB storage for each fine-tuned model.
Compute requirement. It requires more VRAM/training time.
Compositionality. Hard to combine multiple models.

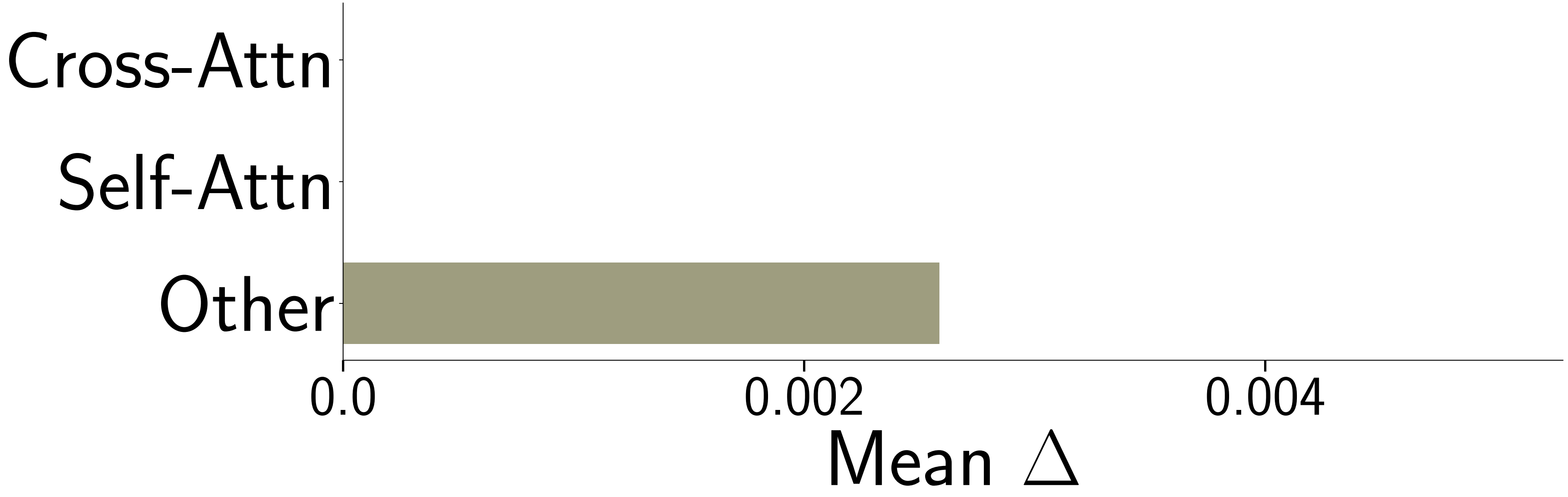
Analyze change in weights

$$\Delta_l = \frac{||\theta_l' - \theta_l||}{||\theta_l||} \quad \text{where} \quad \begin{array}{l} \theta_l' \text{ : updated weights} \\ \theta_l \text{ : pretrained weights} \end{array}$$



Analyze change in weights

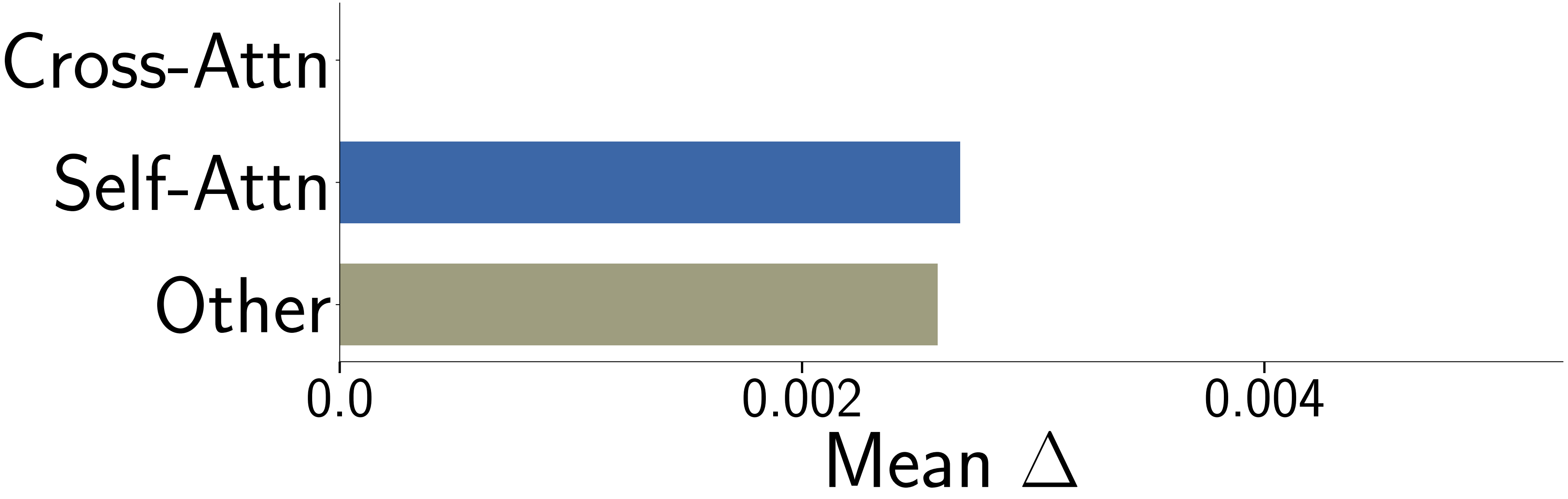
$$\Delta_l = \frac{||\theta'_l - \theta_l||}{||\theta_l||}$$ where θ'_l : updated weights
θ_l : pretrained weights



Analyze change in weights

$$\Delta_l = \frac{||\theta'_l - \theta_l||}{||\theta_l||}$$

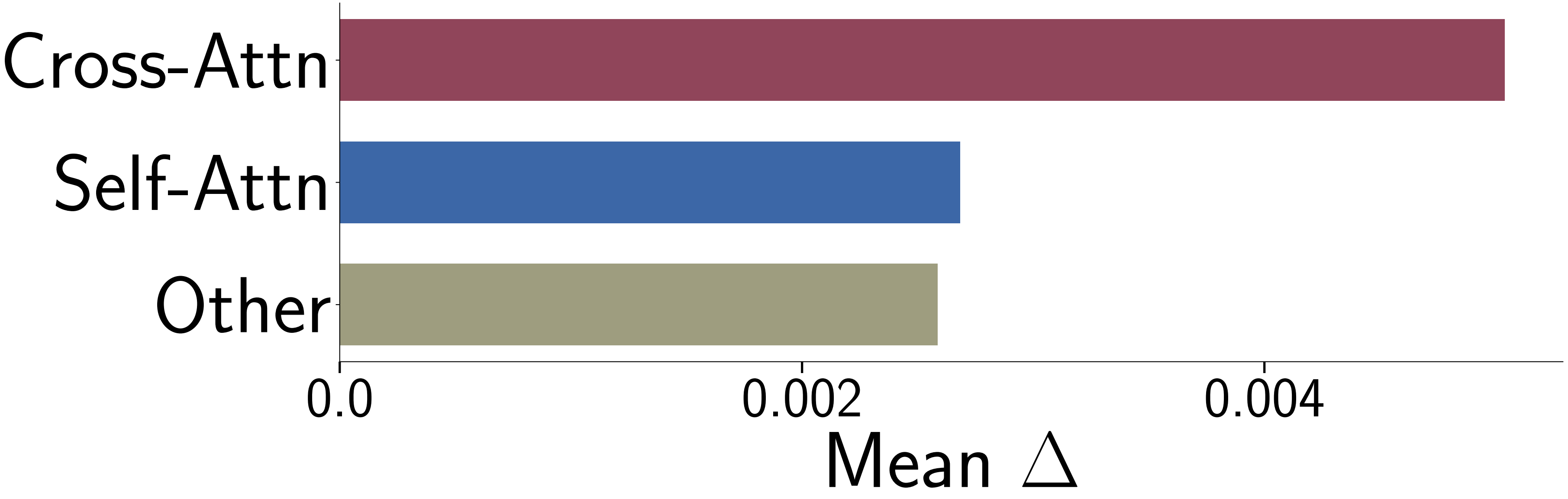
where θ'_l : updated weights
θ_l : pretrained weights



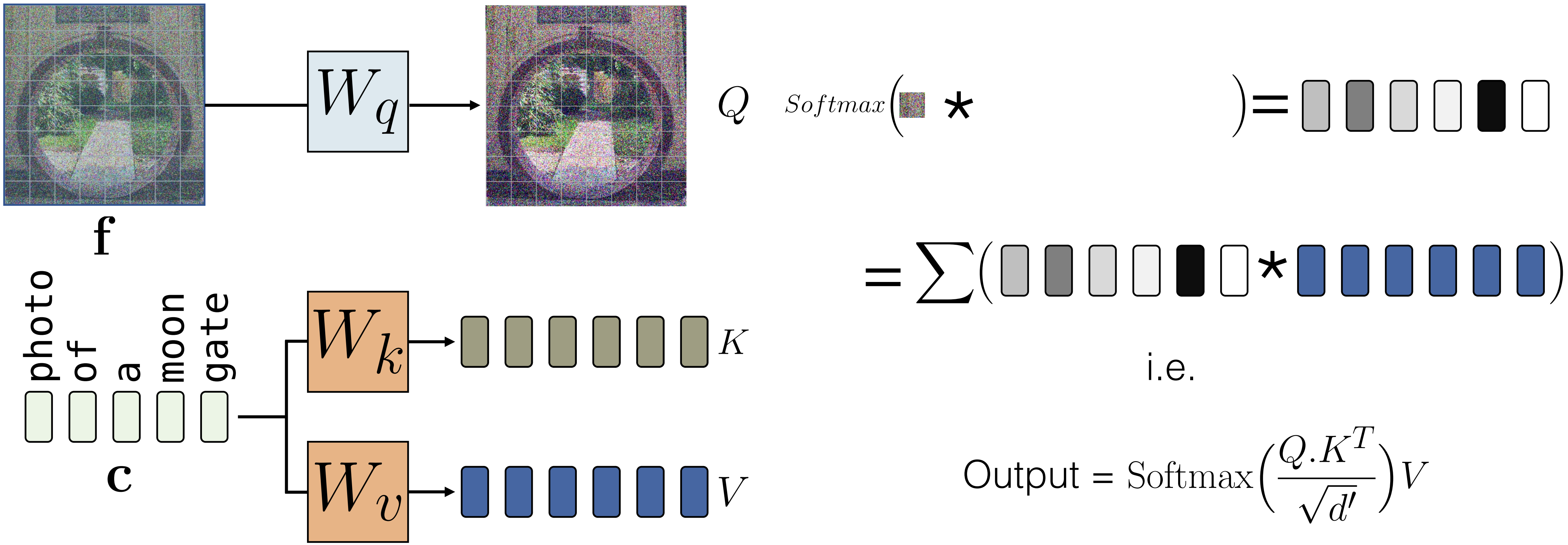
Analyze change in weights

$$\Delta_l = \frac{||\theta'_l - \theta_l||}{||\theta_l||}$$

where θ'_l : updated weights
θ_l : pretrained weights

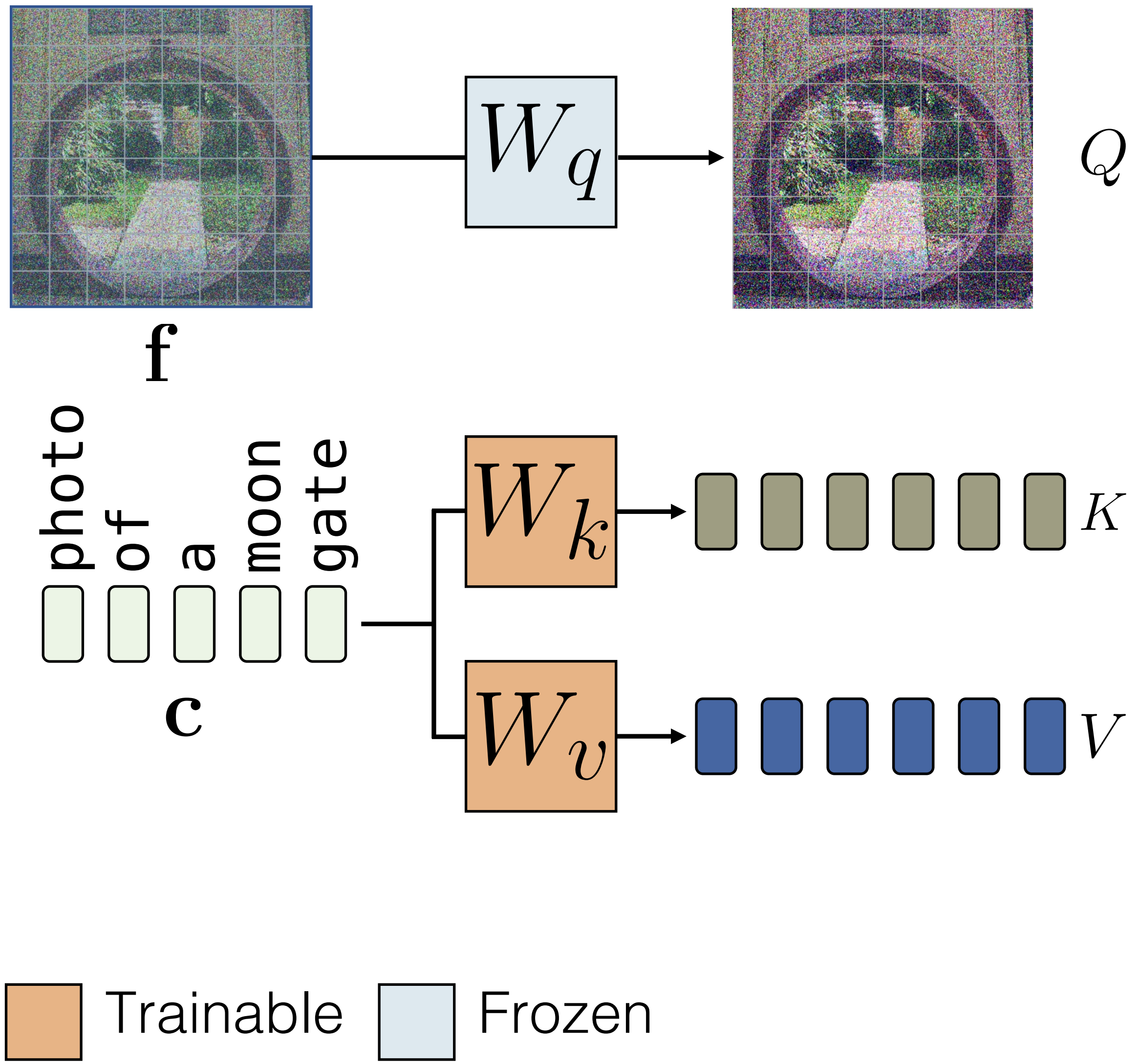


Text-image Cross-Attention



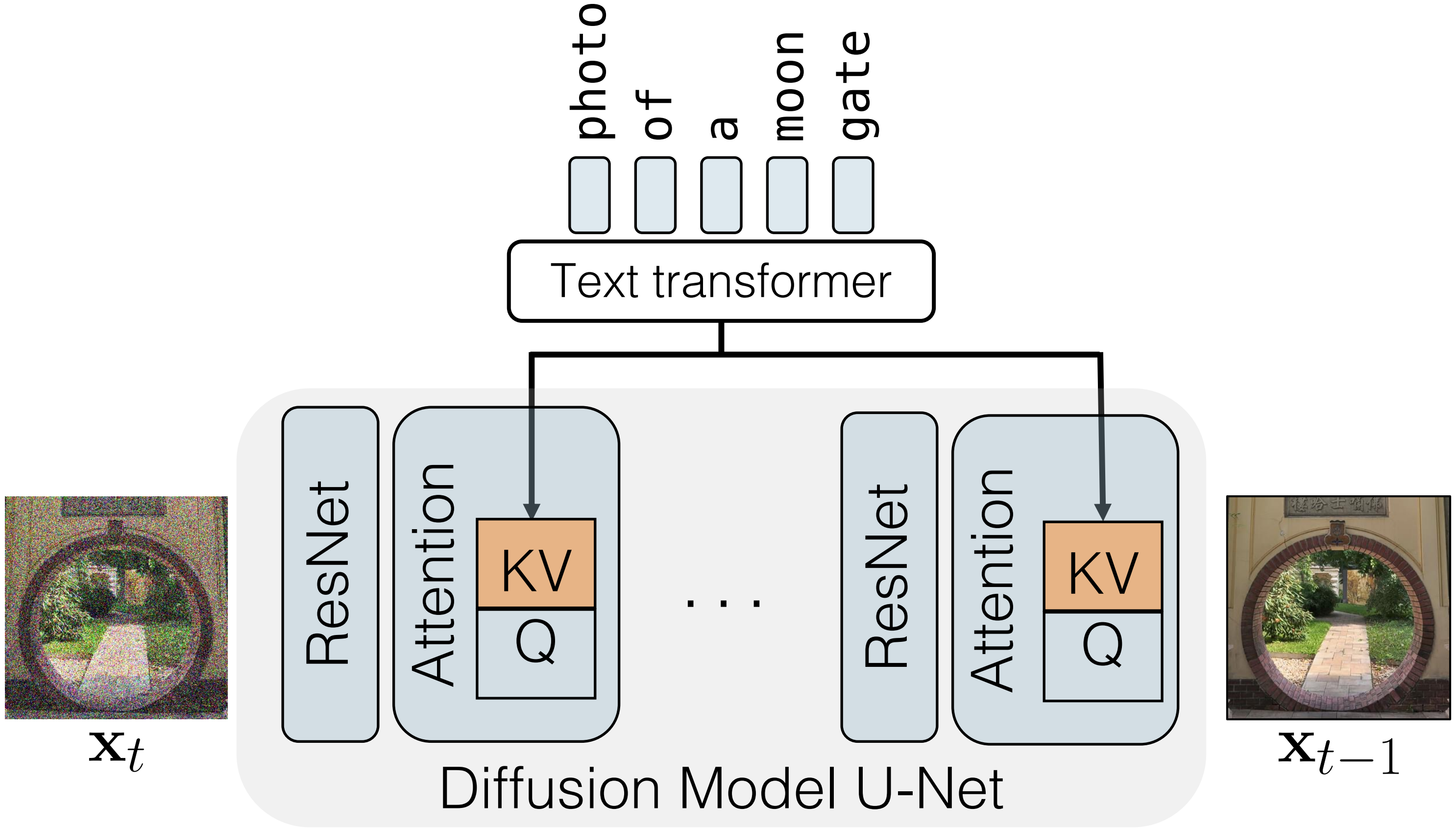
$$\text{Output} = \mathrm{Softmax}\left(\frac{Q.K^T}{\sqrt{d'}}\right)V$$

Text-image Cross-Attention



Text features only input to $\mathrm{W}_{\mathbf{k}}$ and $\mathrm{W}_{\mathbf{v}}$

Only fine-tune cross-attention layers



$$\Delta W_k^*, \Delta W_v^* = \arg \min_{\Delta W_k, \Delta W_v} \mathbb{E}_{\epsilon, \mathbf{x}, \mathbf{c}, t}[||\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)||_2]$$

Trainable Frozen

[Nupur Kumari et al., CVPR 2023]

Generated samples for target concept

`Photo of a moongate`

Pretrained Model

Fine-tuned Model

Generated samples for similar concepts

Photo of a moon

Pretrained Model

Fine-tuned Model

How to prevent overfitting?

Photo of a
{moongate}

Photo of a
{moongate}

...

Target images

+

sky full of stars
and the moon

Blood moon

...

Add regularization images

Generated samples for similar concepts

`Photo of a moon`

Pretrained Model

Fine-tuned Model

Generated samples for similar concepts

Photo of a moon

Pretrained Model

Fine-tuned Model

Personalized concepts

Jun-Yan's dog, Stark

How to describe personalized concepts?

V* dog

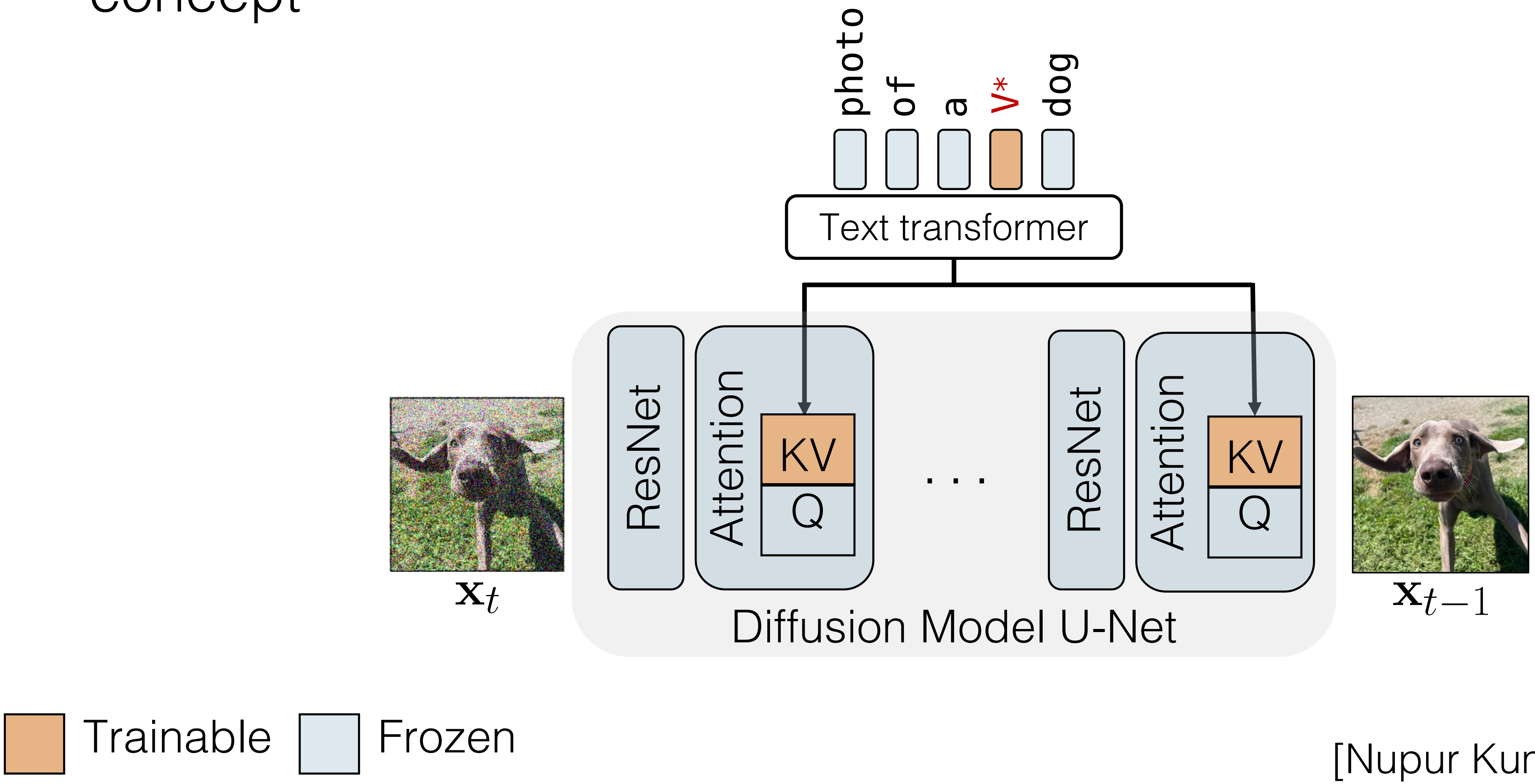
Where V* is a modifier token in the text embedding space

Proposed by Textual Inversion [Rinon Gal et al.]

[Nupur Kumari et al., CVPR 2023]

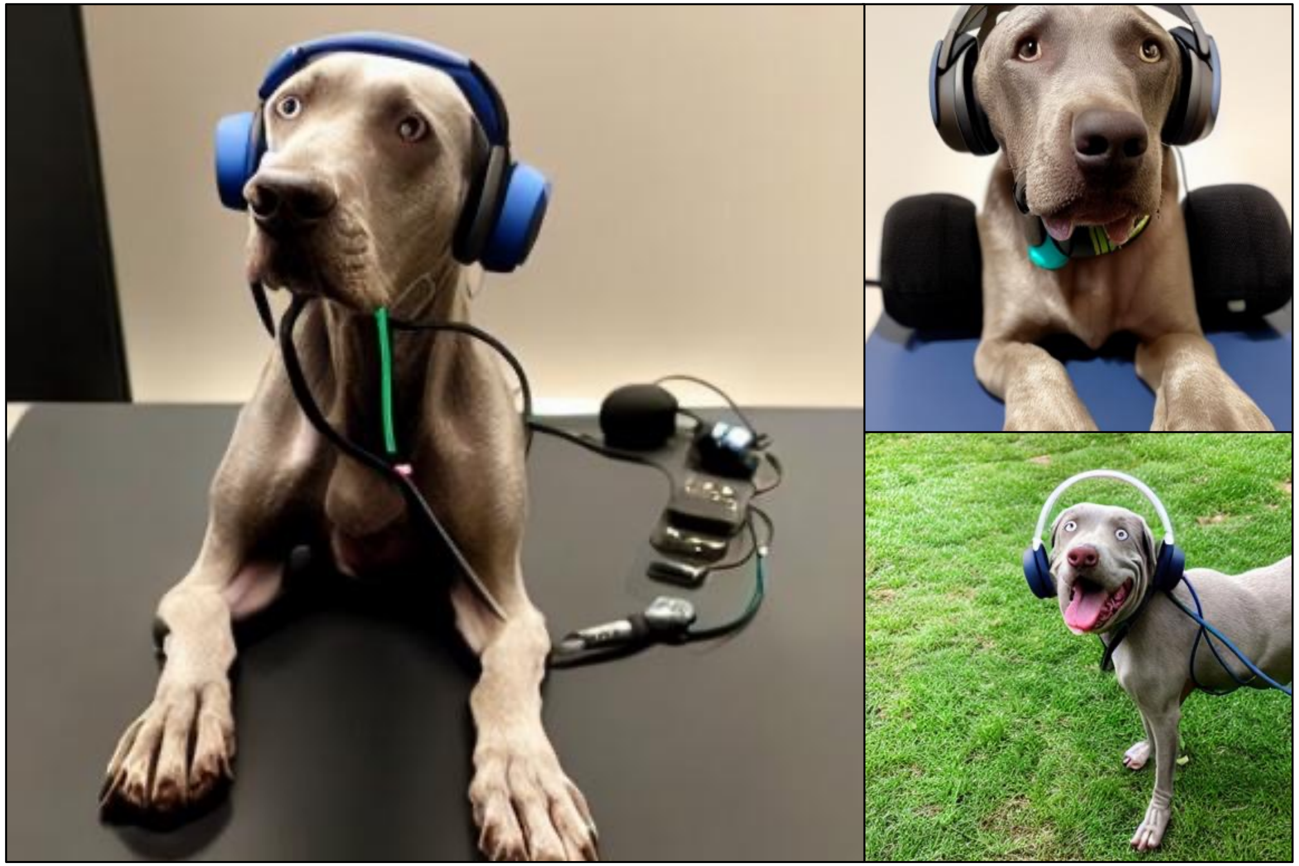
Personalized concepts

Also fine-tune the modifier token V* that describes the personalized concept



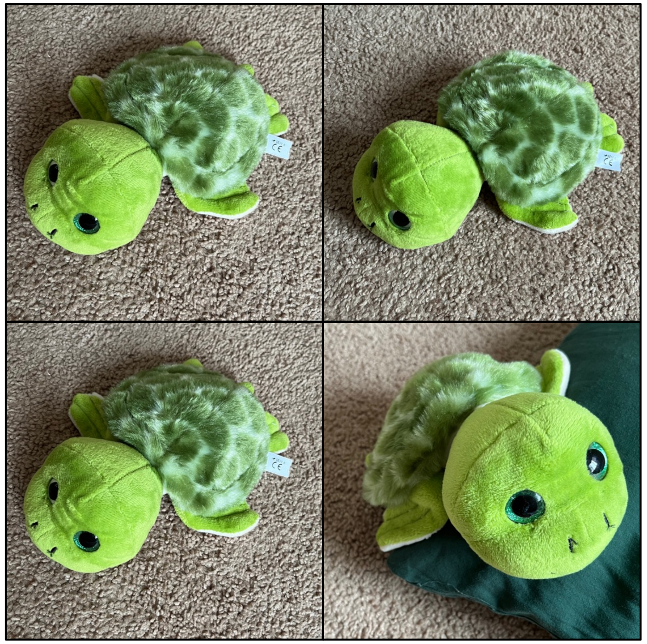
[Nupur Kumari et al., CVPR 2023]

Single concept results



V* dog wearing headphones

Single concept results



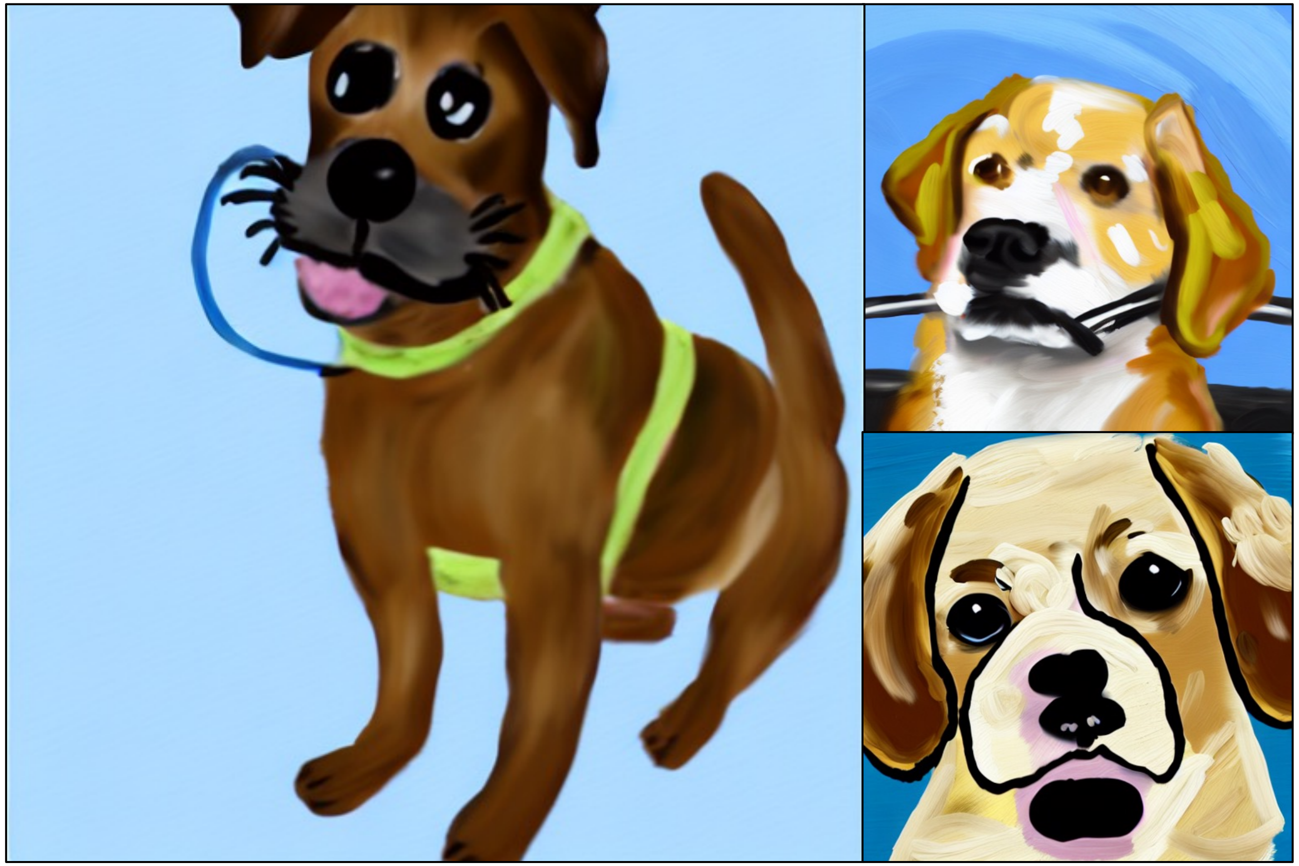
A watercolor painting of V* tortoise plushy on a mountain

Single concept results

V* table and an orange sofa

Results: specific art style

Drawings from Aaron Hertzmann



Painting of dog in the style of V* art

Multiple new concepts?

+

?

Joint training

1. Combine the training dataset of multiple concepts

Target images

V∗ dog

Moongate

Regularization images

Dog Cute dog

Wisdom moon Gated entry

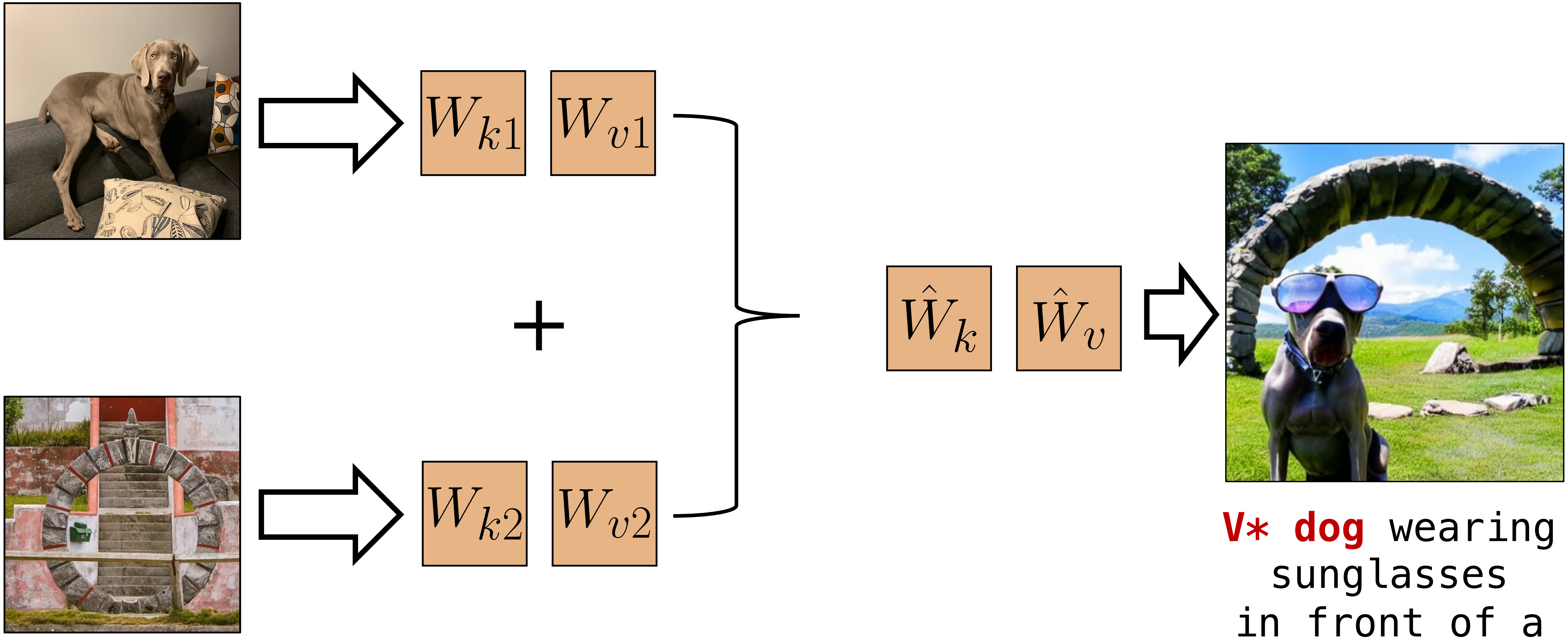
Joint training

Requires re-training for each choice of composition

100 concepts -> 4950 combinations of **two** concepts.

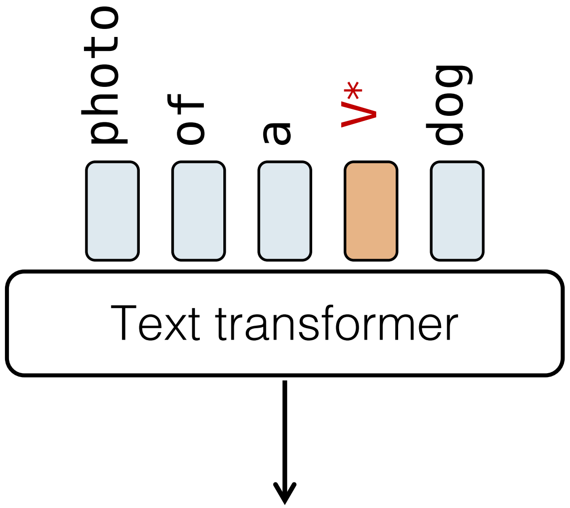
100 concepts -> 161, 700 combinations of **three** concepts.

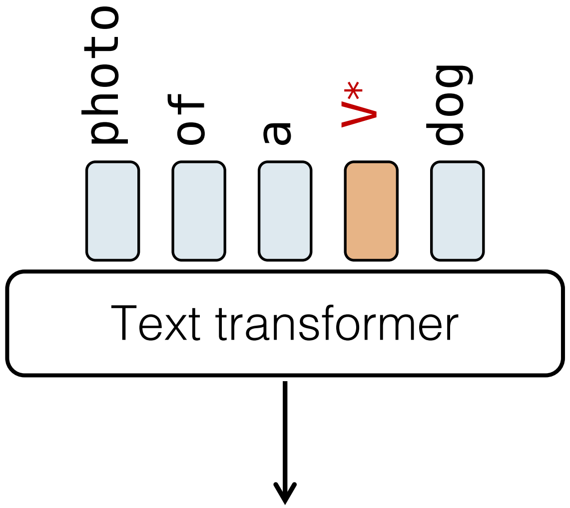
Can we merge weights of individual concepts?



V* dog wearing sunglasses in front of a **moongate**

Objective function for merging weights





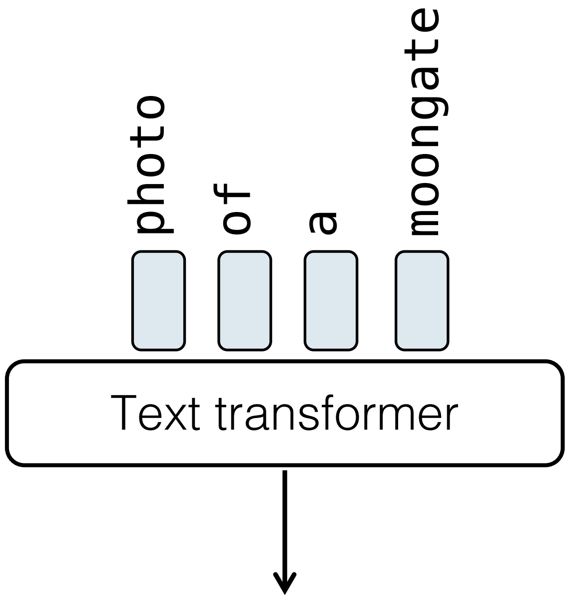
$$W_{k1} * [c_1 \cdots c_k] = \hat{W}_k * [c_1 \cdots c_k]$$

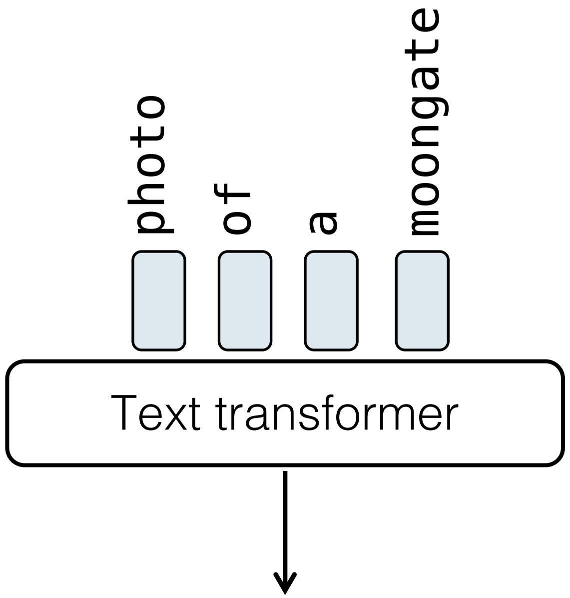
Fine-tuned weights for V* dog

Merged weights

[Nupur Kumari et al., CVPR 2023]

Objective function for merging weights





$$\boxed{W_{k2}} * [c_{k+1} \cdots c_N] = \boxed{\hat{W}_k} * [c_{k+1} \cdots c_N]$$

Fine-tuned weights for moongate

Merged weights

[Nupur Kumari et al., CVPR 2023]

Constrained least square problem

Stay close to pretrained weights W_0 for random text prompts C_{reg}.

$$\hat{W} = \arg\min_{W} ||WC_{\text{reg}}^{\top} - W_0 C_{\text{reg}}^{\top}||_F$$

s.t. $\hat{W}[c_1 \cdots c_N] = [W_1 c_1 \cdots W_2 c_N]$

C: target prompts, e.g., {`photo of a V* dog, photo of moongate`}

[Nupur Kumari et al., CVPR 2023]

Constrained least square problem

Constrained least square problem

$$\hat{W} = \arg\min_{W} ||WC_{\mathbf{reg}}^{\top} - W_0 C_{\mathbf{reg}}^{\top}||_F$$

s.t. $\hat{W}[c_1 \cdots c_N] = [W_1 c_1 \cdots W_2 c_N]$

[Nupur Kumari et al., CVPR 2023]

Constrained least square problem

Constrained least square problem

$$\hat{W} = \arg\min_{W} ||W C_{\text{reg}}^{\top} - W_0 C_{\text{reg}}^{\top}||_F$$

s.t. $\hat{W}[c_1 \cdots c_N] = [W_1 c_1 \cdots W_2 c_N]$

Close-form solution for solving for W and v,

$$\hat{W} = W_0 + \mathbf{v}^{\top}\mathbf{d}, \text{ where } \mathbf{d} = C(C_{\text{reg}}^{\top} C_{\text{reg}})^{-1}$$
$$\text{and } \mathbf{v}^{\top} = (V - W_0 C^{\top})(\mathbf{d} C^{\top})^{-1}$$

[Nupur Kumari et al., CVPR 2023]

Two concept results

V_1^* dog in front of moongate

Two concept results

The V_1^* cat is sitting inside a V_2^* wooden pot and looking up

Two concept results

V_1* chair with the V_2* cat sitting on it near a beach

Two concept results

V_1^* flower in the V_2^* wooden pot on a table

Two concept results

V_1* art style painting of V_2* wooden pot

Drawings from Aaron Hertzmann

Qualitative comparison (single-concept)

Target Images

V* teddybear in
Times Square??

Qualitative comparison (single-concept)

Target Images

Custom Diffusion (Ours)

DreamBooth

Textual Inversion

`V* teddybear in Times Square`

Qualitative comparison (multi-concept)

Target Images

Custom Diffusion (Ours)

DreamBooth

Textual Inversion

V_1* flower in the V_2* wooden pot on a table

Limitations

Ours

V_1* dog and a V_2* cat playing together

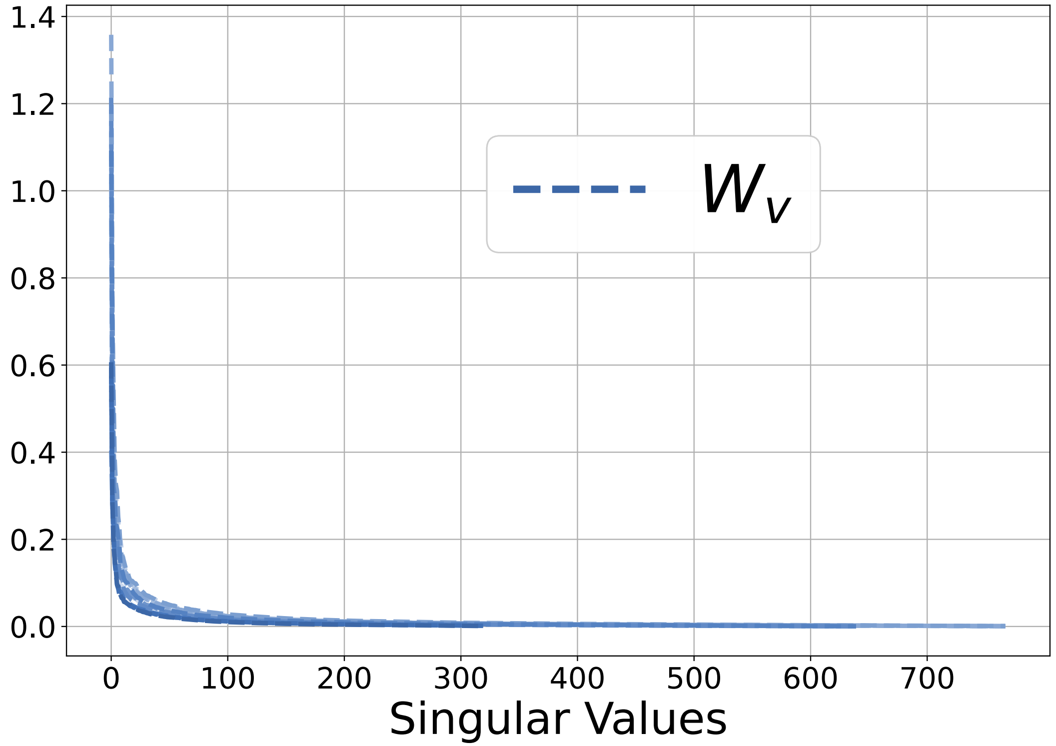
Pretrained model

dog and a cat playing together

Memory requirement

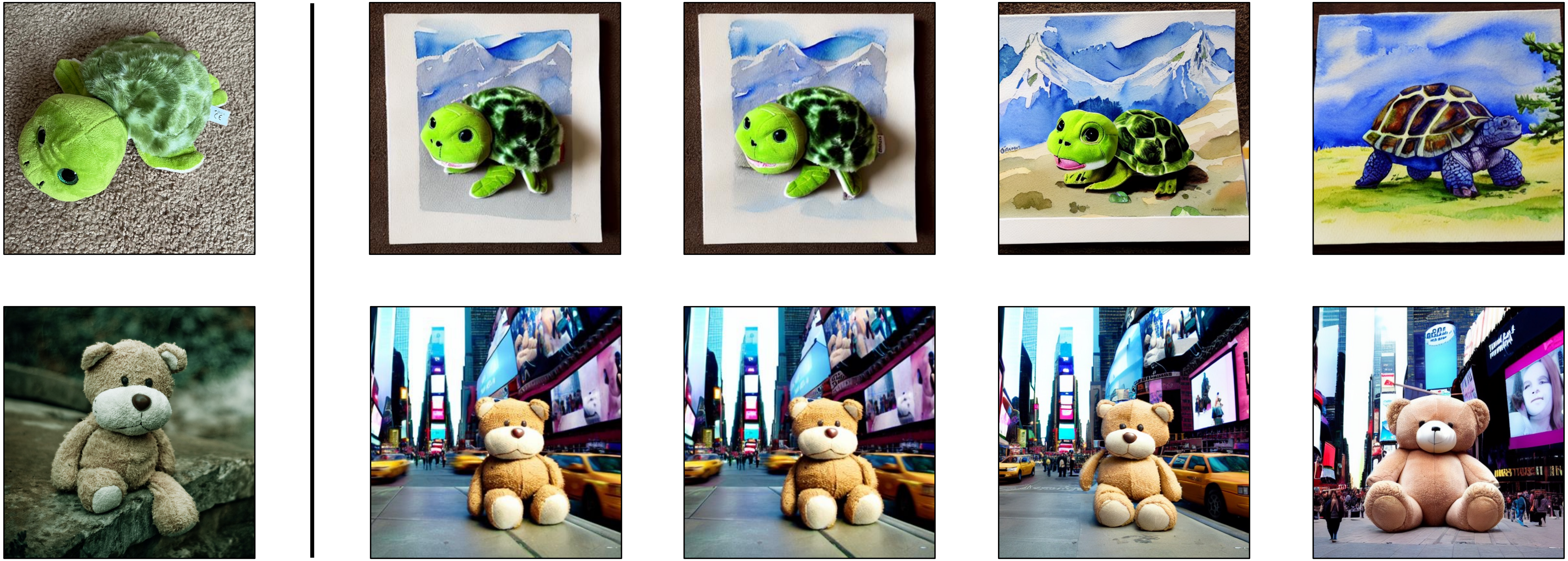
Each custom diffusion model: 75MB storage

Analyze the difference in pretrained and fine-tuned weights



Compressing fine-tuned weights

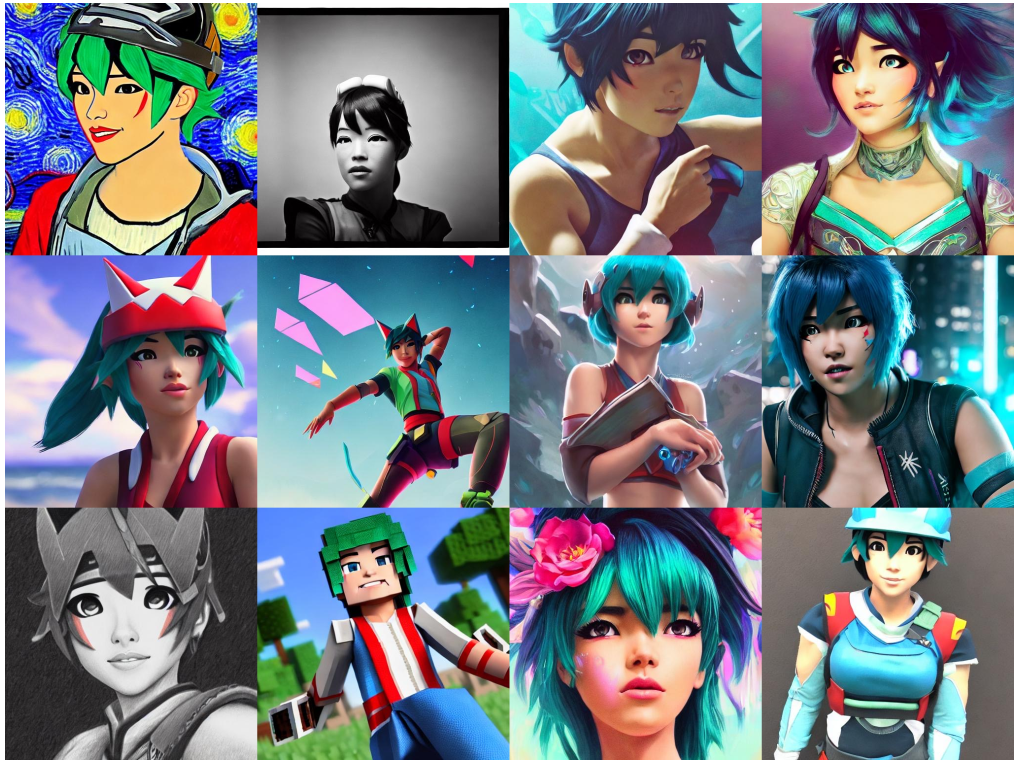
75MB | 15MB | 0.1MB | 0.08MB

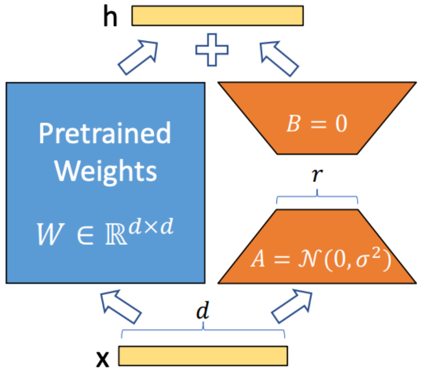


Target image | Custom Diffusion | Top 20% rank | 1 Rank | 0 Rank

Low-rank Adaptation (Lora)

- Lora: Low-rank adaptation of large language models





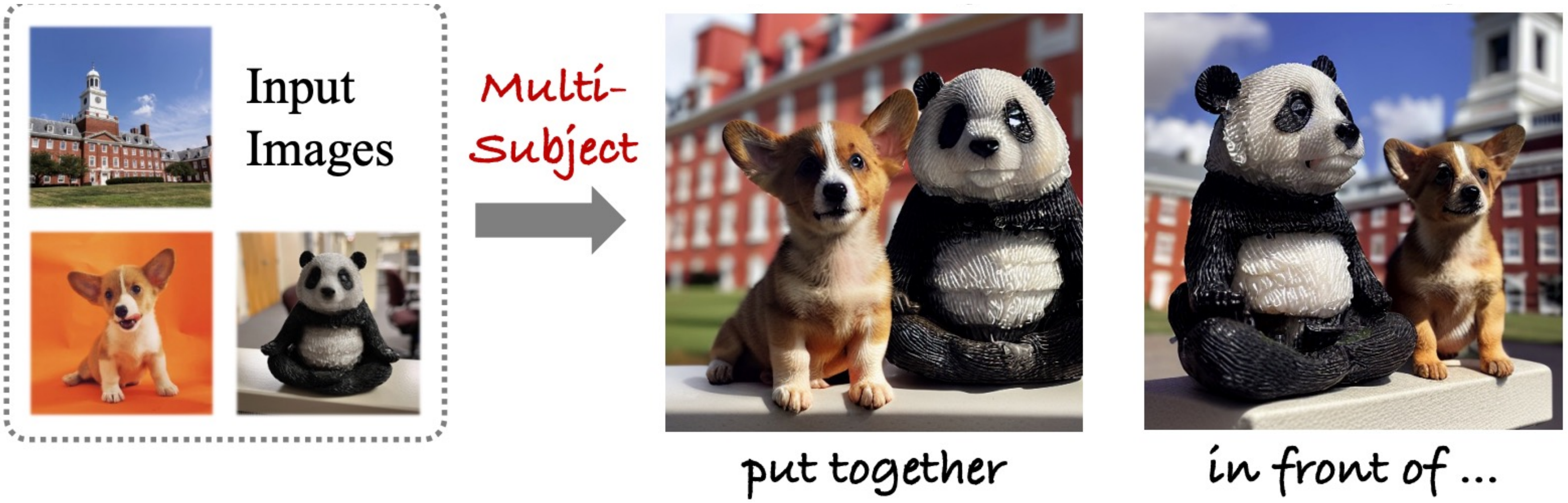
Original weights

$$W = W_0 + BA$$

Low-rank difference

Lora [Edward J. Hu*, Yelong Shen*, et al., ICLR 2022]
Lora + Dreambooth (by Simo Ryu): https://github.com/cloneofsimo/lora

Low-rank Adaptation (SVDiff)



- Composing multiple concepts

$$\Sigma_{\boldsymbol{\delta}'} = \mathrm{diag}(\mathrm{ReLU}(\boldsymbol{\sigma} + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2))$$

SVDiff [Han et all., ICLR 2022]

Low-rank Adaptation (Rank-1)

- Rank-1 Model Editing
- Used in GAN fine-tuning [Bau et al., 2020] and LLM factual editing [Meng et al., 2022]

$$\hat{W} = W + \Lambda(C^{-1}\boldsymbol{i}_*)^T.$$

$$\Lambda \ = \ (\boldsymbol{o}_* \ - \ W\boldsymbol{i}_*)/[(\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*)]$$

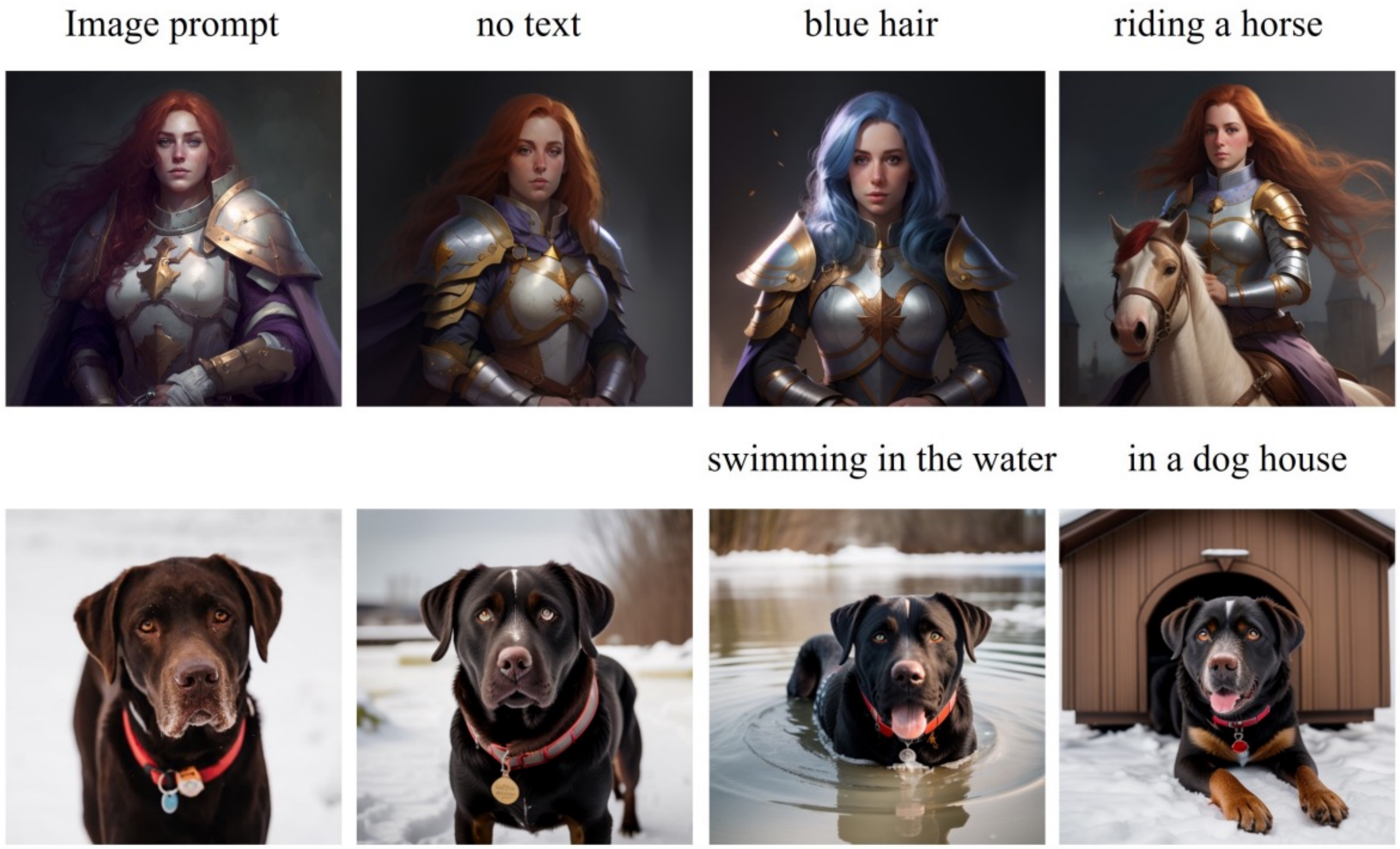
Please see their paper for more details including key lock

Perfusion [Tewel et all., SIGGRAPH 2023]

Optimization is too Slow!

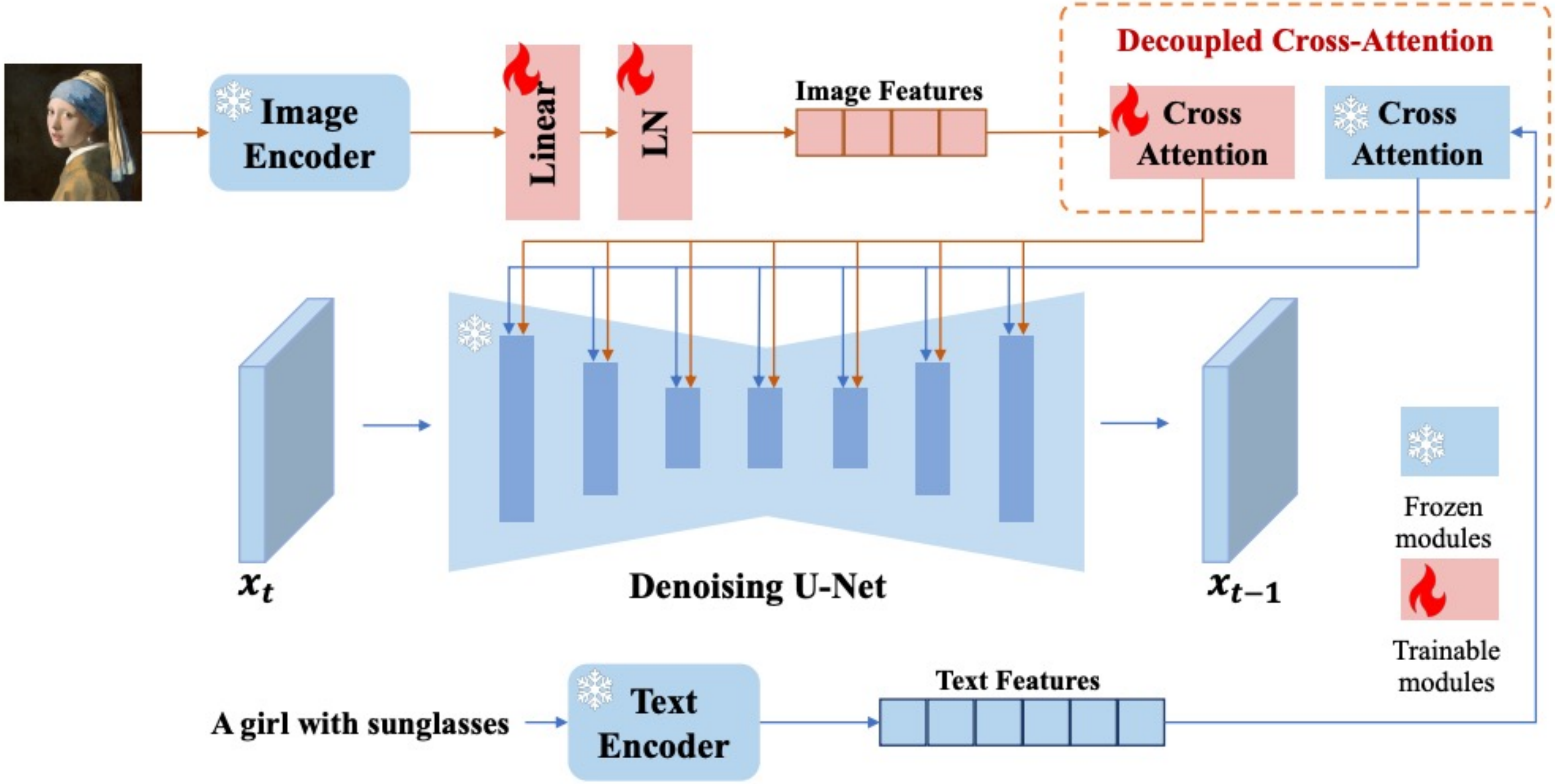
Encoder-based Methods

Image Prompt Adapter (IP-Adapter)



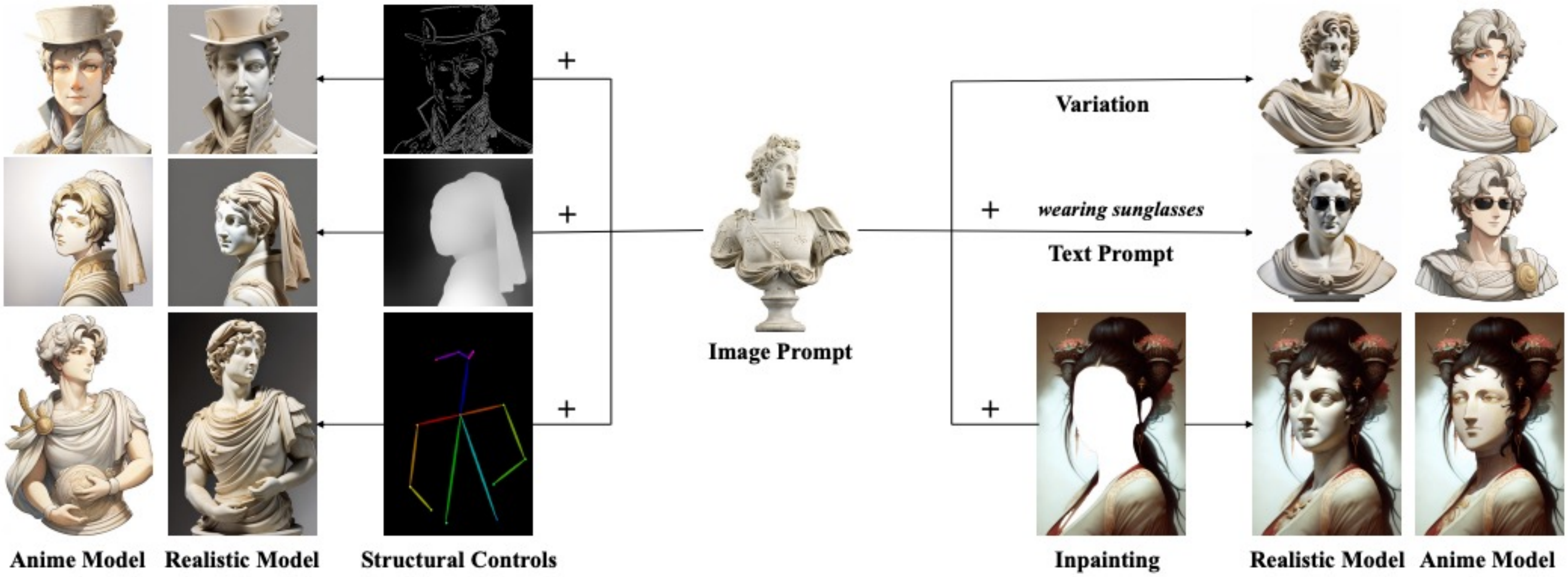
[He Yu et al., CVPR 2024]

Image Prompt Adapter (IP-Adapter)



[He Yu et al., CVPR 2024]

Image Prompt Adapter (IP-Adapter)



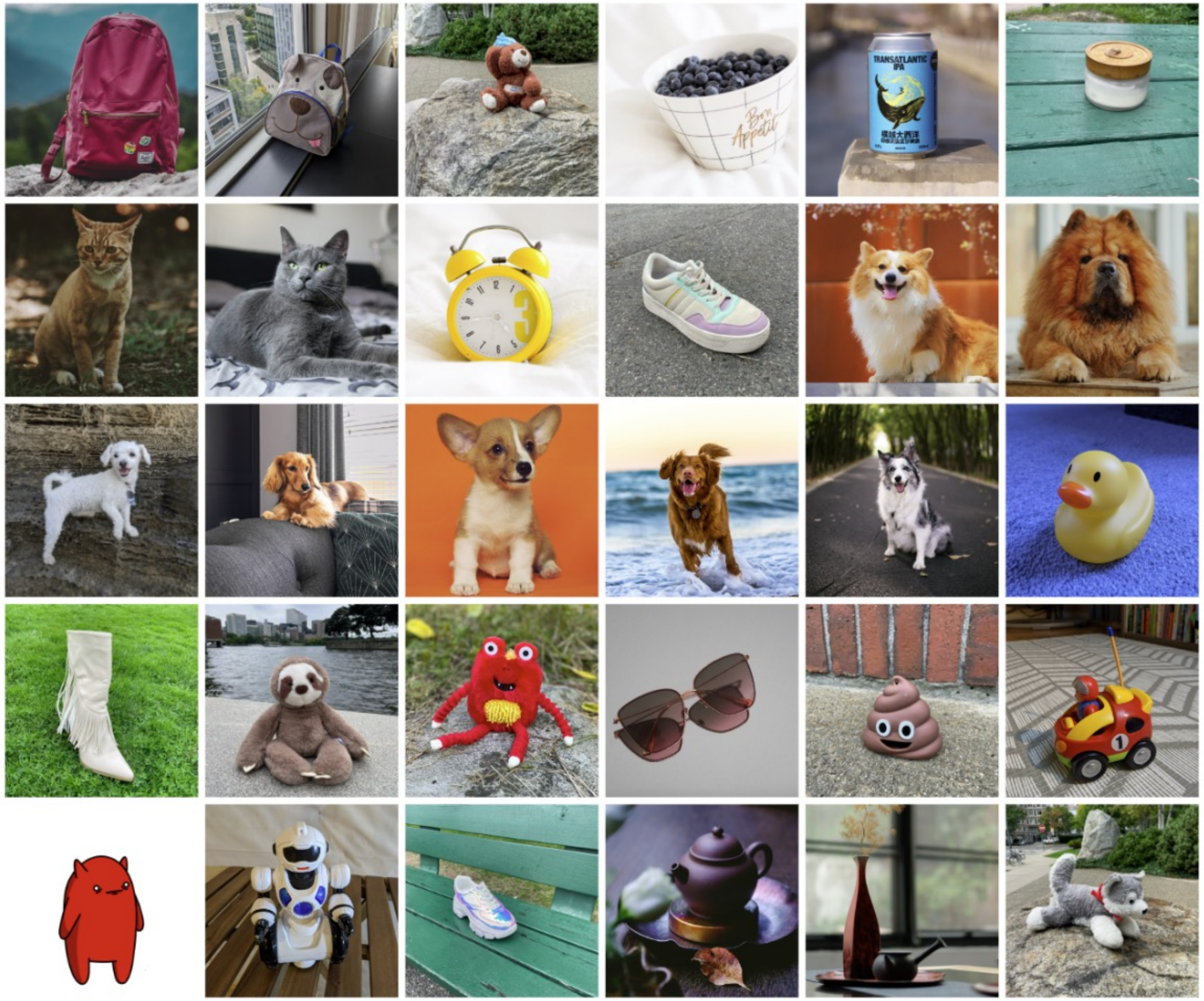
[He Yu et al., CVPR 2024]

Optimization + encoder (5-15 steps)

[Rinon Gal et al., arXiv 2023]

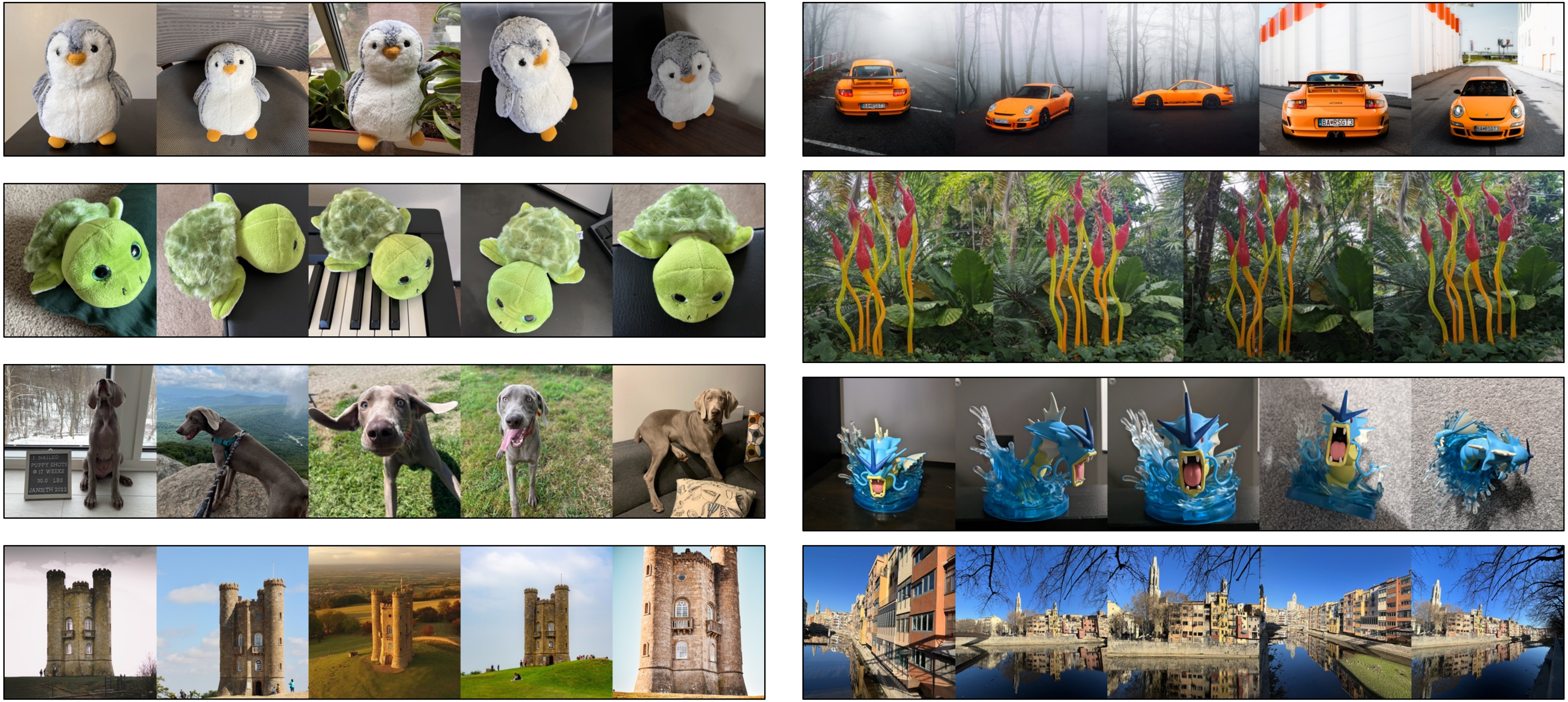
Datasets

DreamBooth Dataset: 30 subjects



[Nataniel Ruiz et al., CVPR 2023]

CustomConcept101: 101 concepts



[Nupur Kumari et al., CVPR 2023]