STRUCT Group Paper Reading

MoMA: Multimodal LLM Adapter for Fast Personalized Image Generation

Kunpeng Song[1,2], Yizhe zhu[1], Bingchen Liu[1], Qing Yan[1], Ahmed Elgammal[2], and Xiao Yang[1]

[1] ByteDance
[2] Rutgers University

2024/5/19

Outline

1 Background

2 Author

3 Method

4 Experiments

Background

Customized image generation

Background

Personalized image generation

Background

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

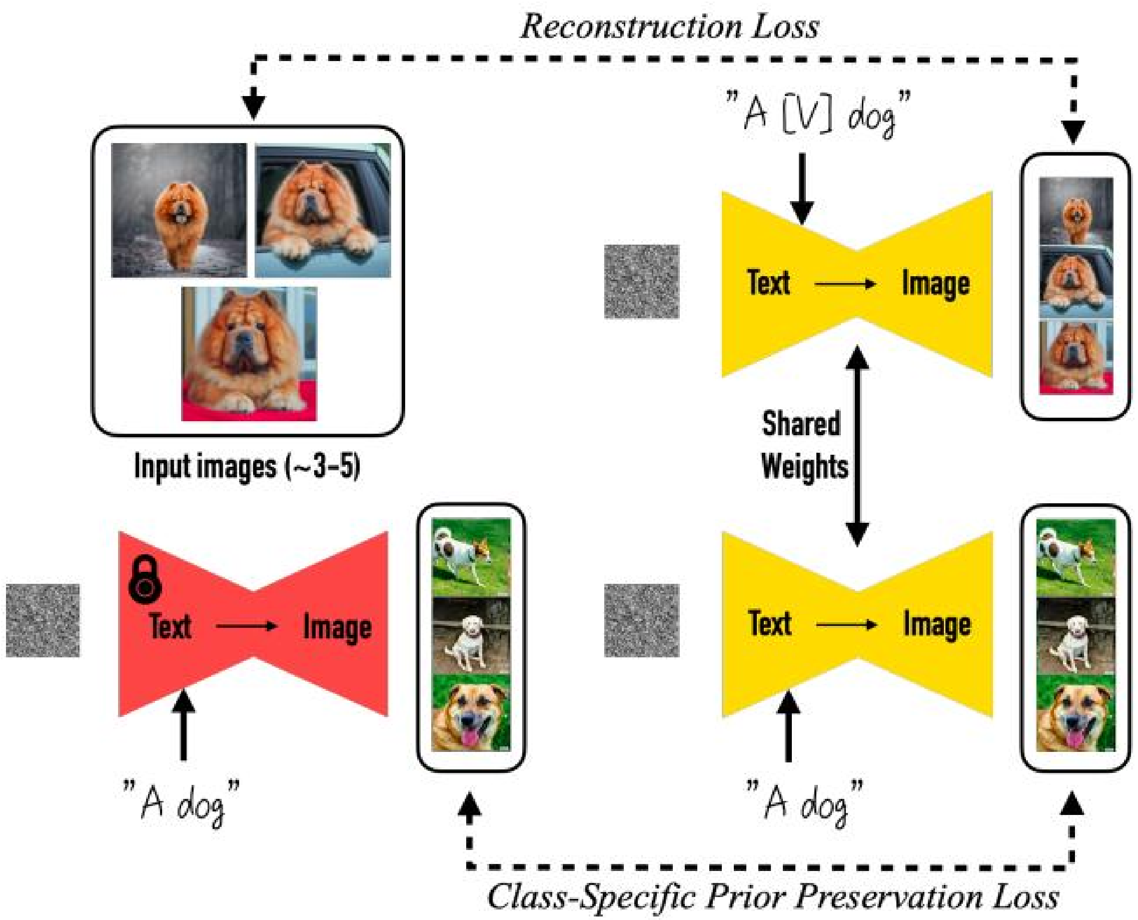
Nataniel Ruiz[*,1,2] Yuanzhen Li[1] Varun Jampani[1]
Yael Pritch[1] Michael Rubinstein[1] Kfir Aberman[1]
[1] Google Research [2] Boston University

Input images | in the Acropolis | in a doghouse | in a bucket | getting a haircut

Ruiz N, Li Y, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22500-22510.

Background

- **DreamBooth**
  - **Finetune T2I model with unique identifier**
  - **Regularize the model with class-specific prior**



Loss function:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+ \lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2],$$

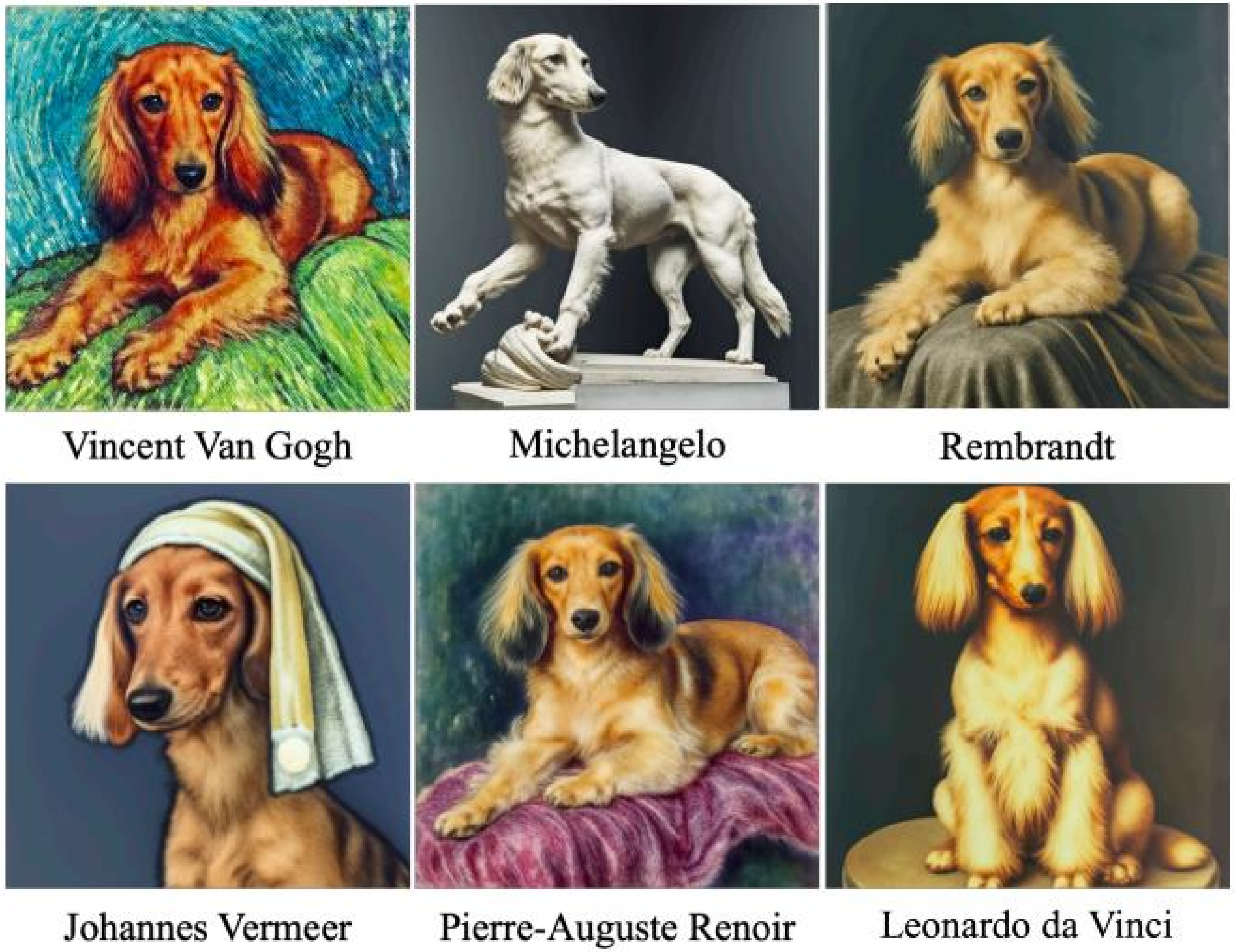
Background

Background editing

Background

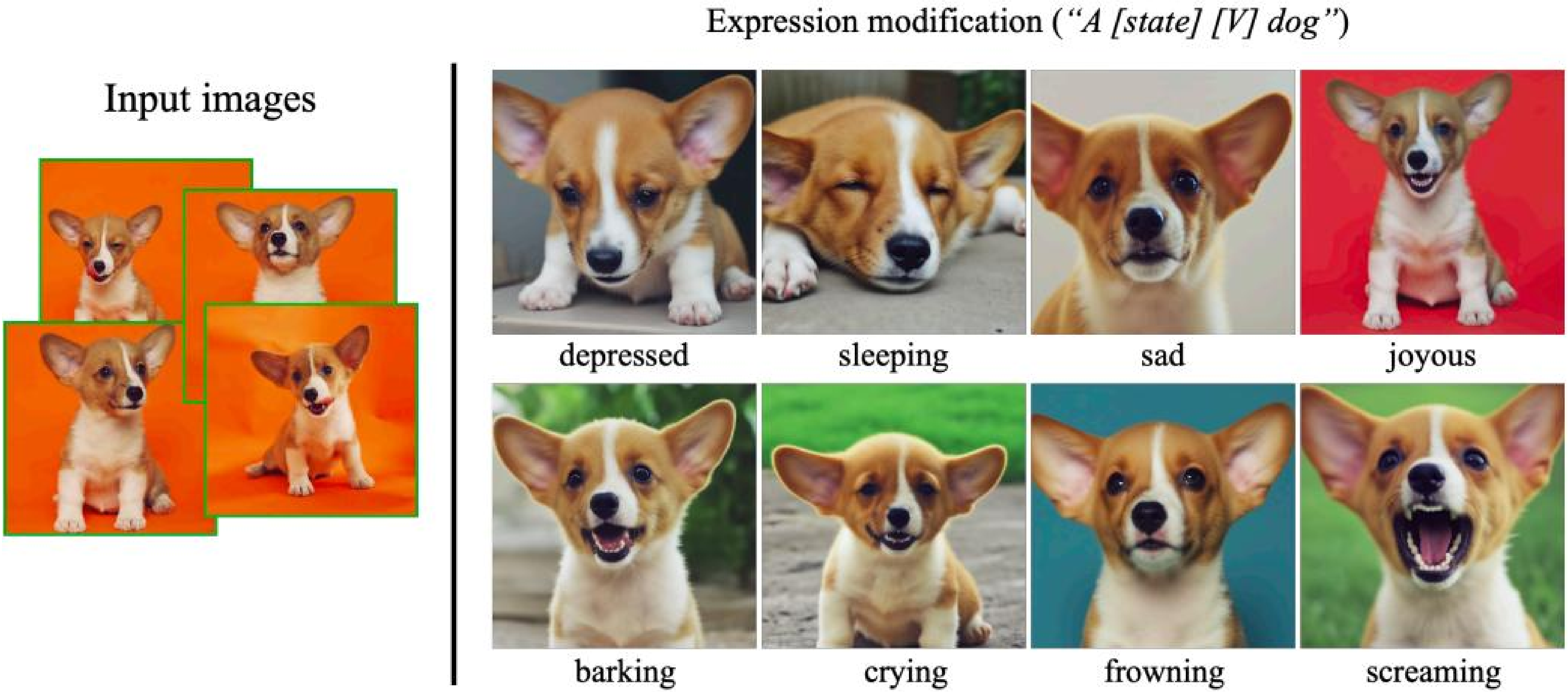
Style editing

Input images



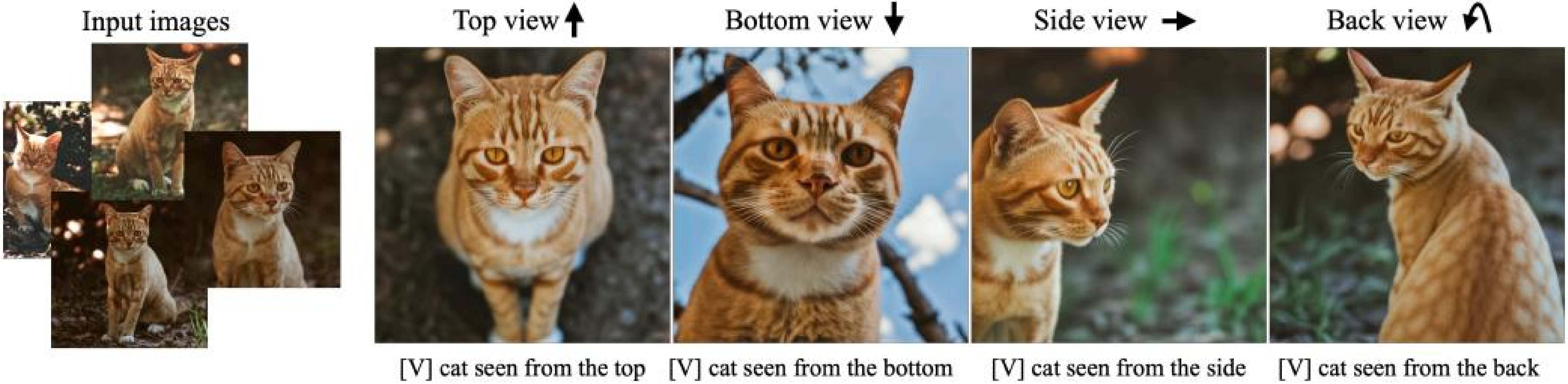
Background

Expression editing



Background

View editing



Background

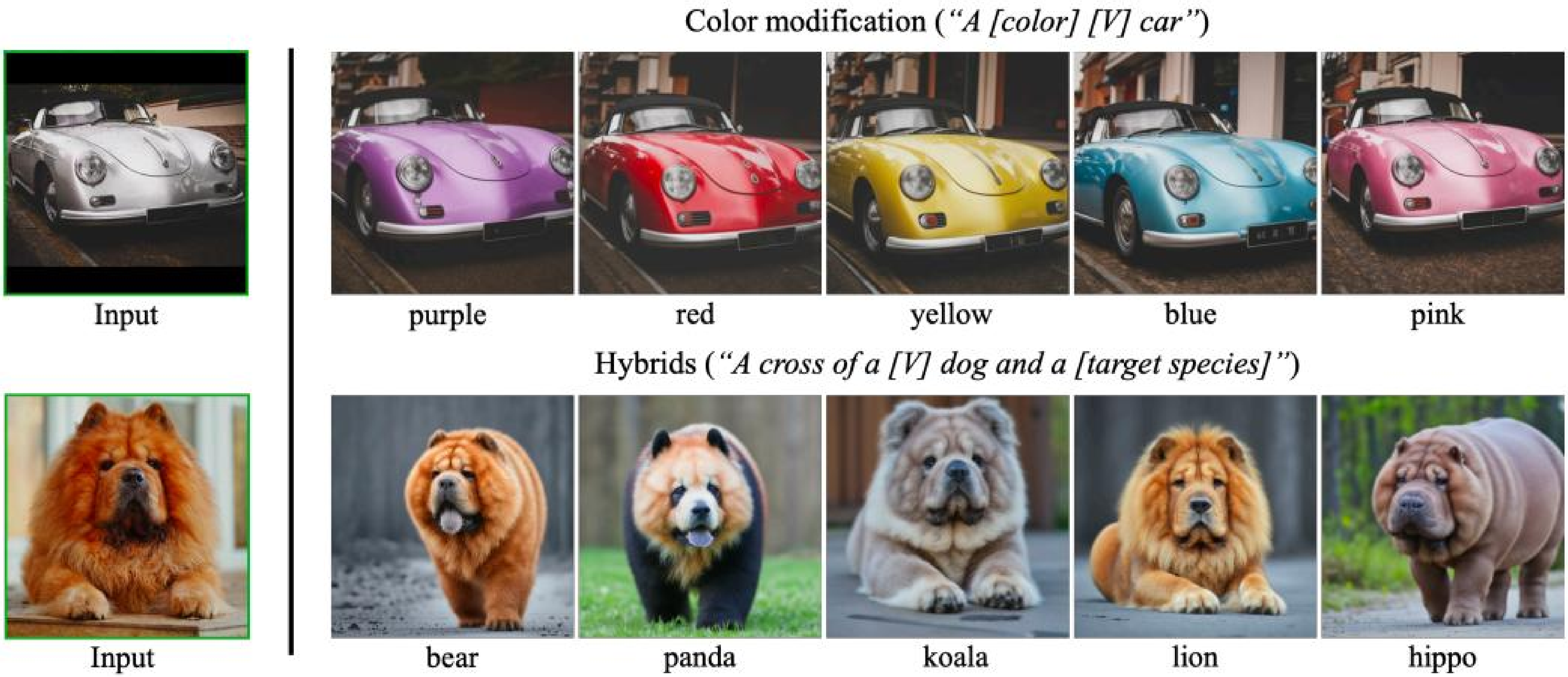
Accessary editing

Input images

a [V] dog wearing a police/chef/witch outfit

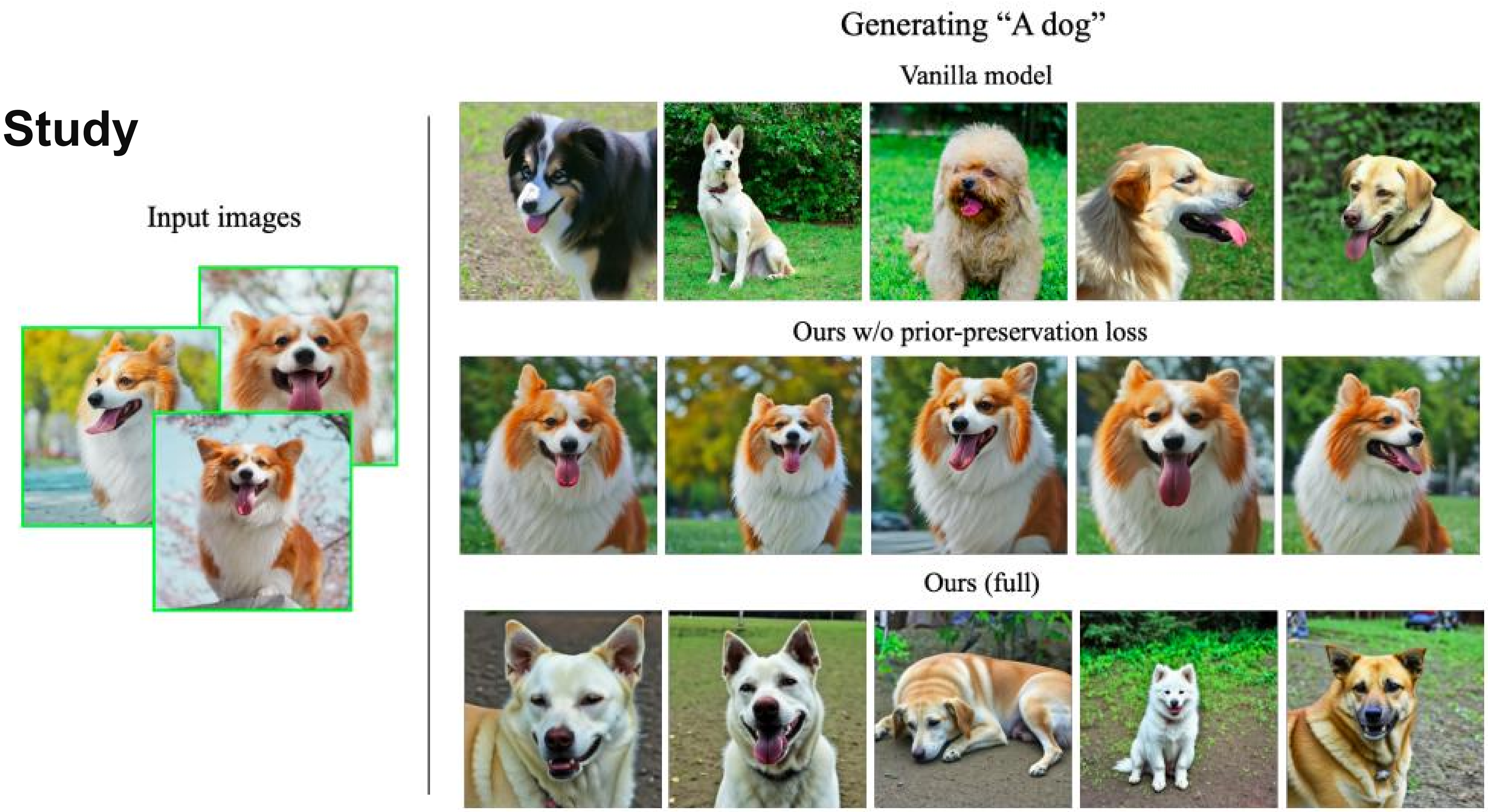
Background

Color editing & attribute editing



Background

Ablation Study



Background

Background

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

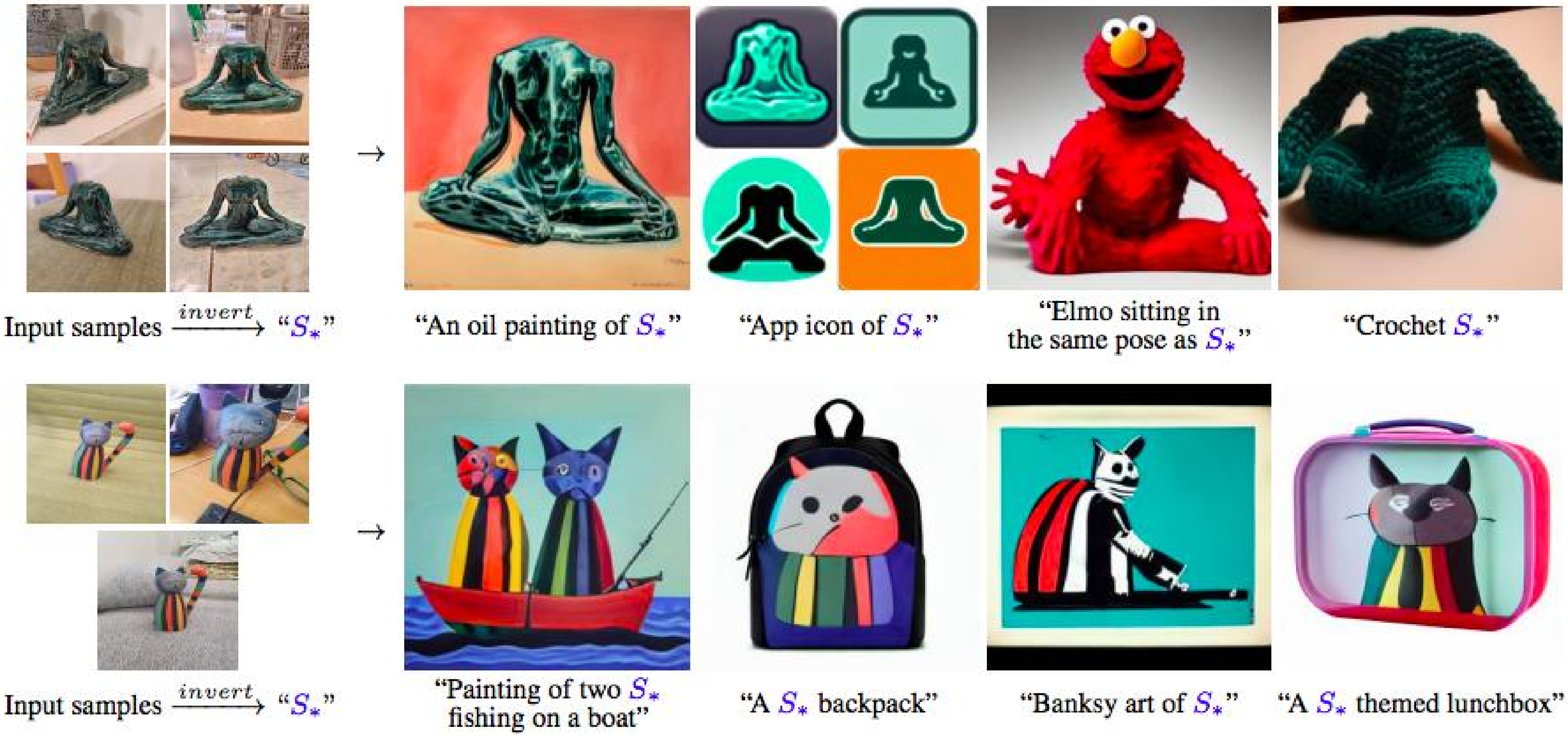
Rinon Gal[1,2*] Yuval Alaluf[1] Yuval Atzmon[2] Or Patashnik[1]

Amit H. Bermano[1] Gal Chechik[2] Daniel Cohen-Or[1]

[1]Tel-Aviv University [2]NVIDIA

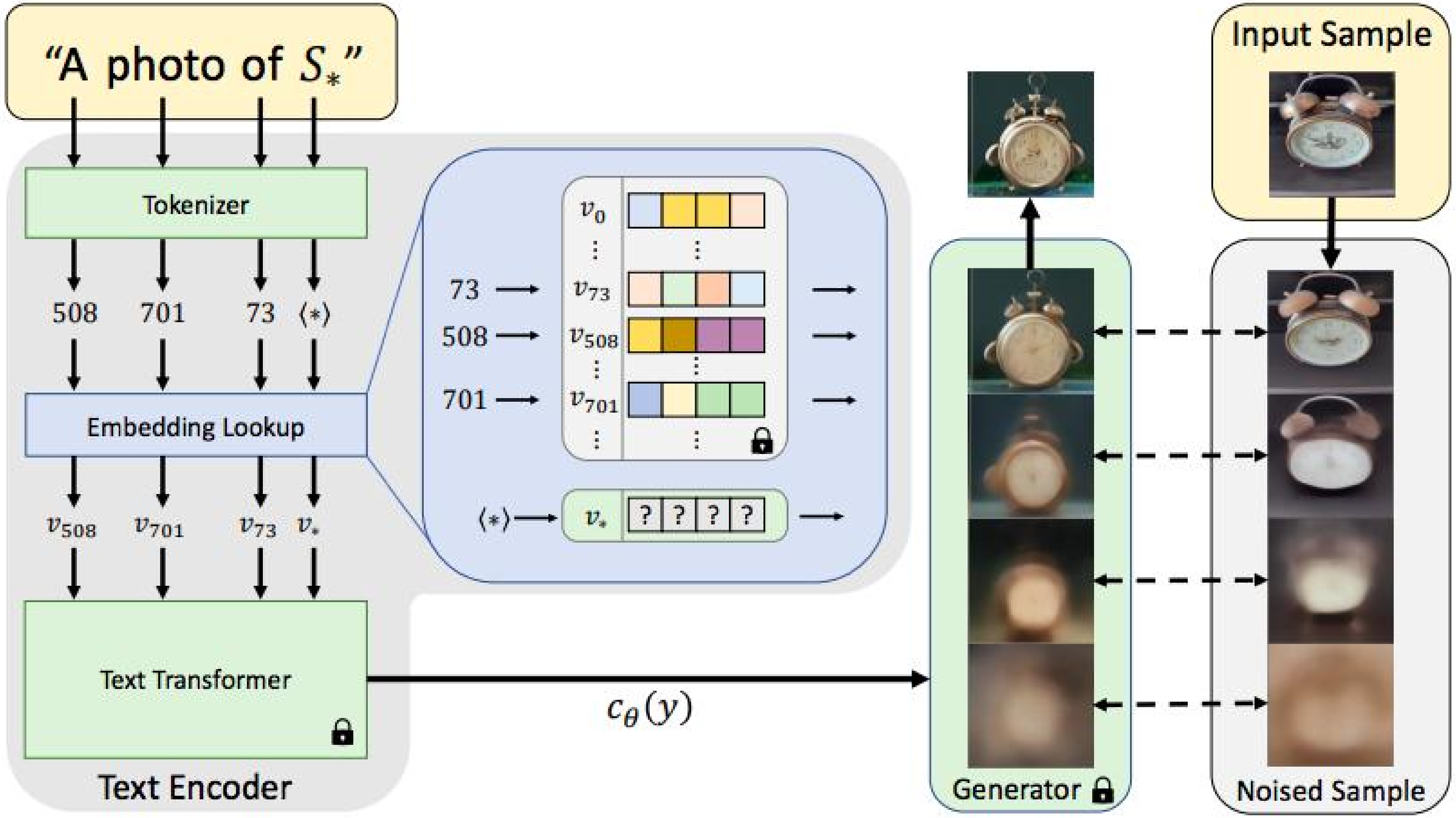
Gal R, Alaluf Y, Atzmon Y, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion[J]. arXiv preprint arXiv:2208.01618, 2022.

Background


Input samples $\xrightarrow{invert}$ "S_*"
"An oil painting of S_*"
"App icon of S_*"
"Elmo sitting in the same pose as S_*"
"Crochet S_*"
Input samples $\xrightarrow{invert}$ "S_*"
"Painting of two S_* fishing on a boat"
"A S_* backpack"
"Banksy art of S_*"
"A S_* themed lunchbox"

Background

- **Method overview**



$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$$

Background

Application in style transfer

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_*"

Background

Multi-text inversion

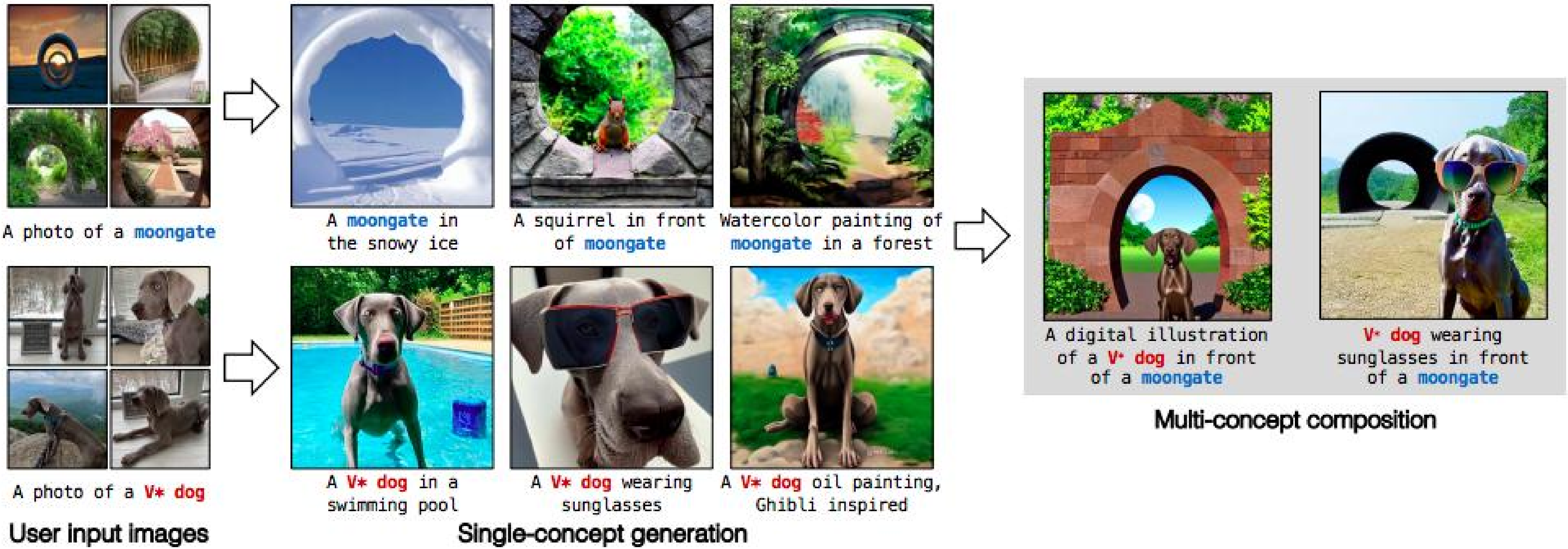
Background

Multi-Concept Customization of Text-to-Image Diffusion

Nupur Kumari[1] Bingliang Zhang[2] Richard Zhang[3] Eli Shechtman[3] Jun-Yan Zhu[1]
[1]Carnegie Mellon University [2]Tsinghua University [3]Adobe Research

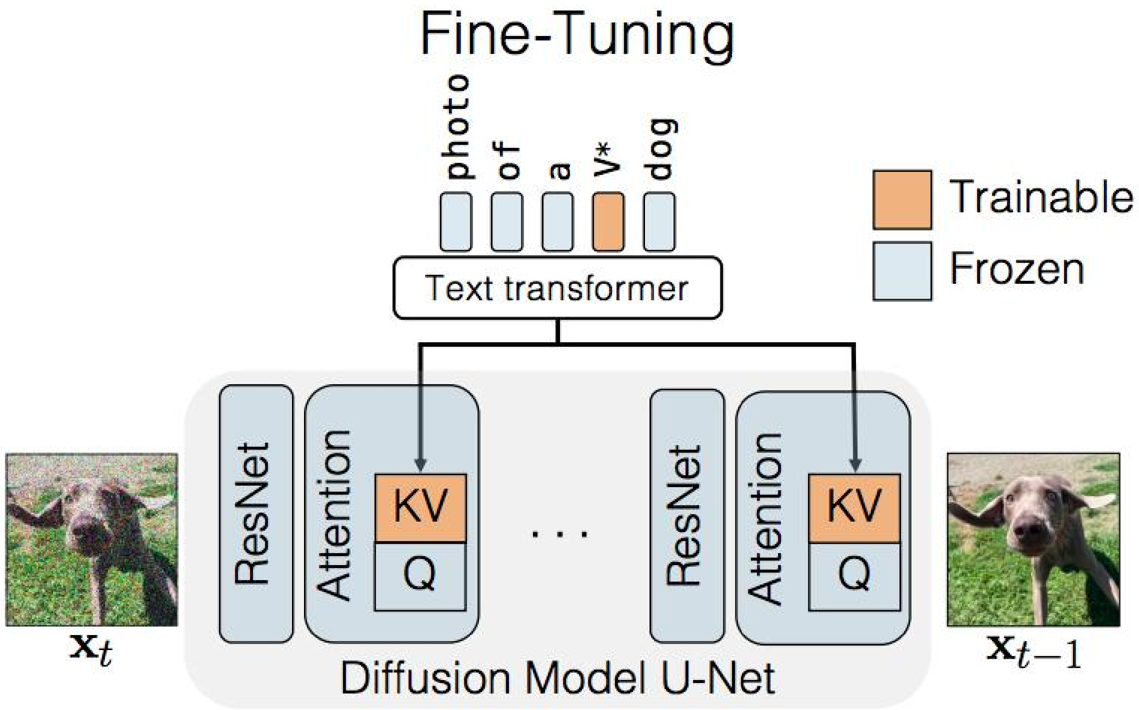
Kumari N, Zhang B, Zhang R, et al. Multi-concept customization of text-to-image diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1931-1941.

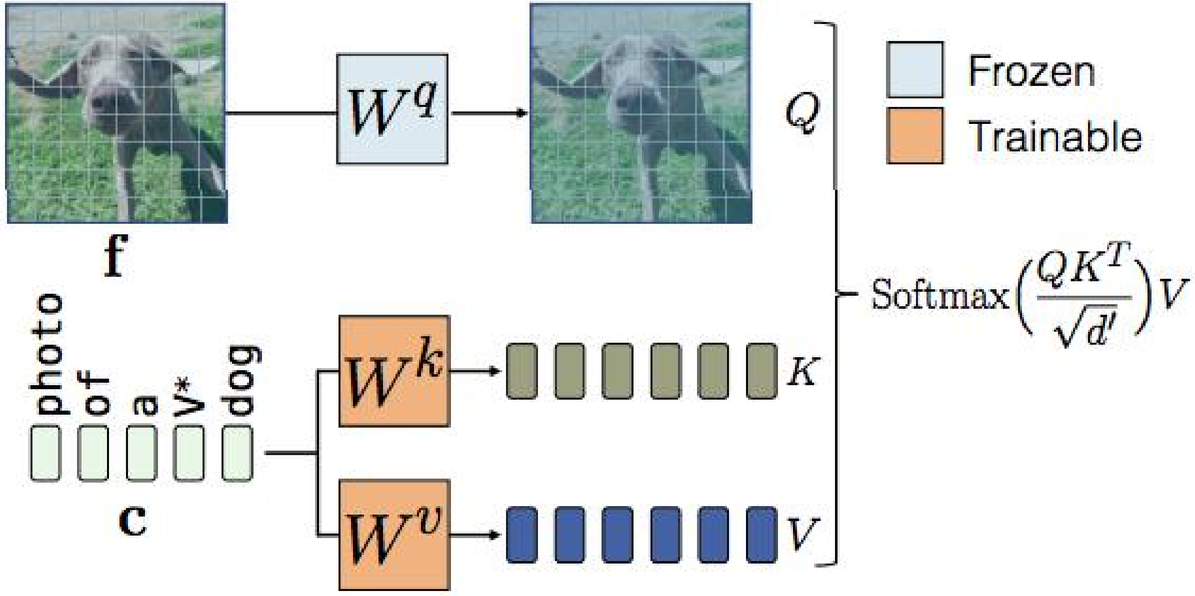
Background


A photo of a moongate
A moongate in the snowy ice
A squirrel in front of moongate
Watercolor painting of moongate in a forest
A photo of a V* dog
A V* dog in a swimming pool
A V* dog wearing sunglasses
A V* dog oil painting, Ghibli inspired
User input images
Single-concept generation
A digital illustration of a V* dog in front of a moongate
V* dog wearing sunglasses in front of a moongate
Multi-concept composition

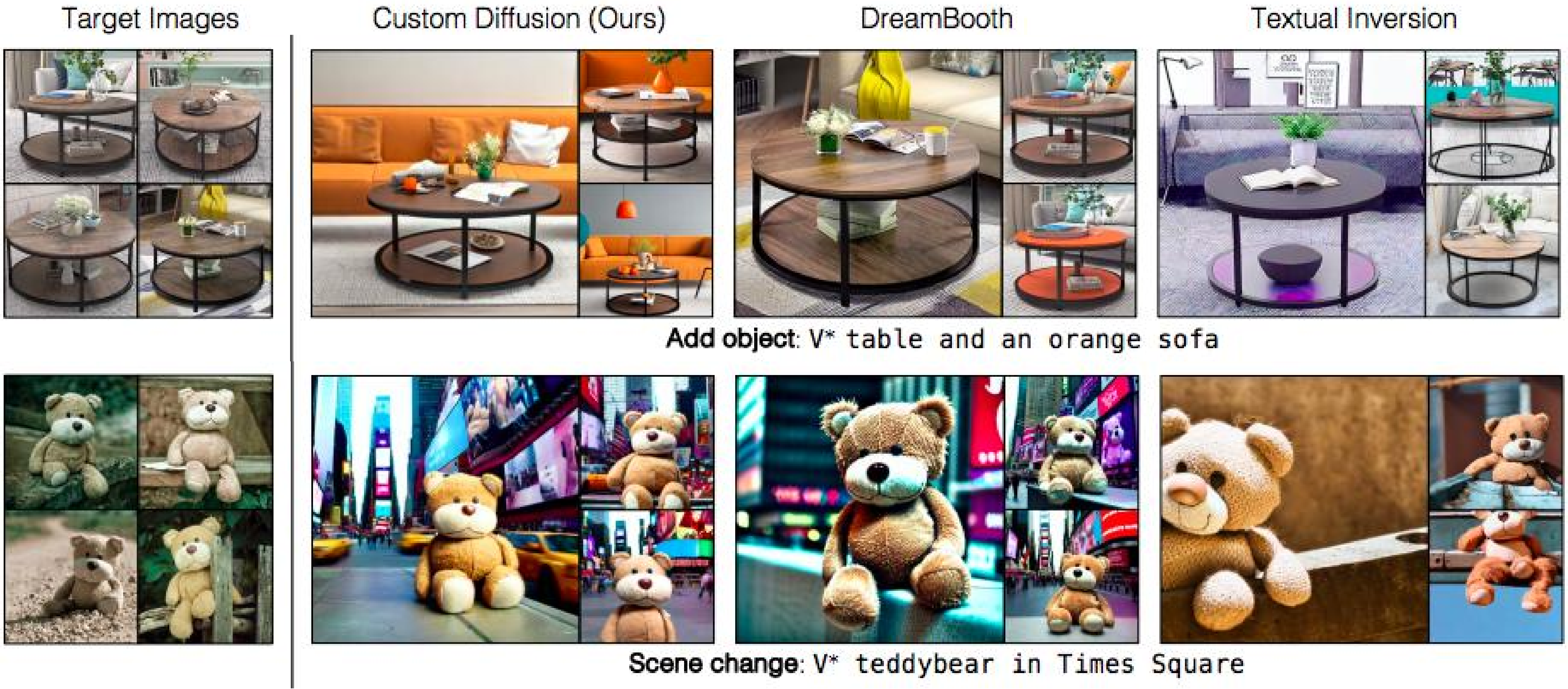
Background

Method overview





Background


Target Images
Custom Diffusion (Ours)
DreamBooth
Textual Inversion
Add object: V* table and an orange sofa
Scene change: V* teddybear in Times Square

Background

- **Multi-concept composition**
  - **Joint training on multiple concepts**
  - **Constrained optimization to merge concepts**

Optimization target

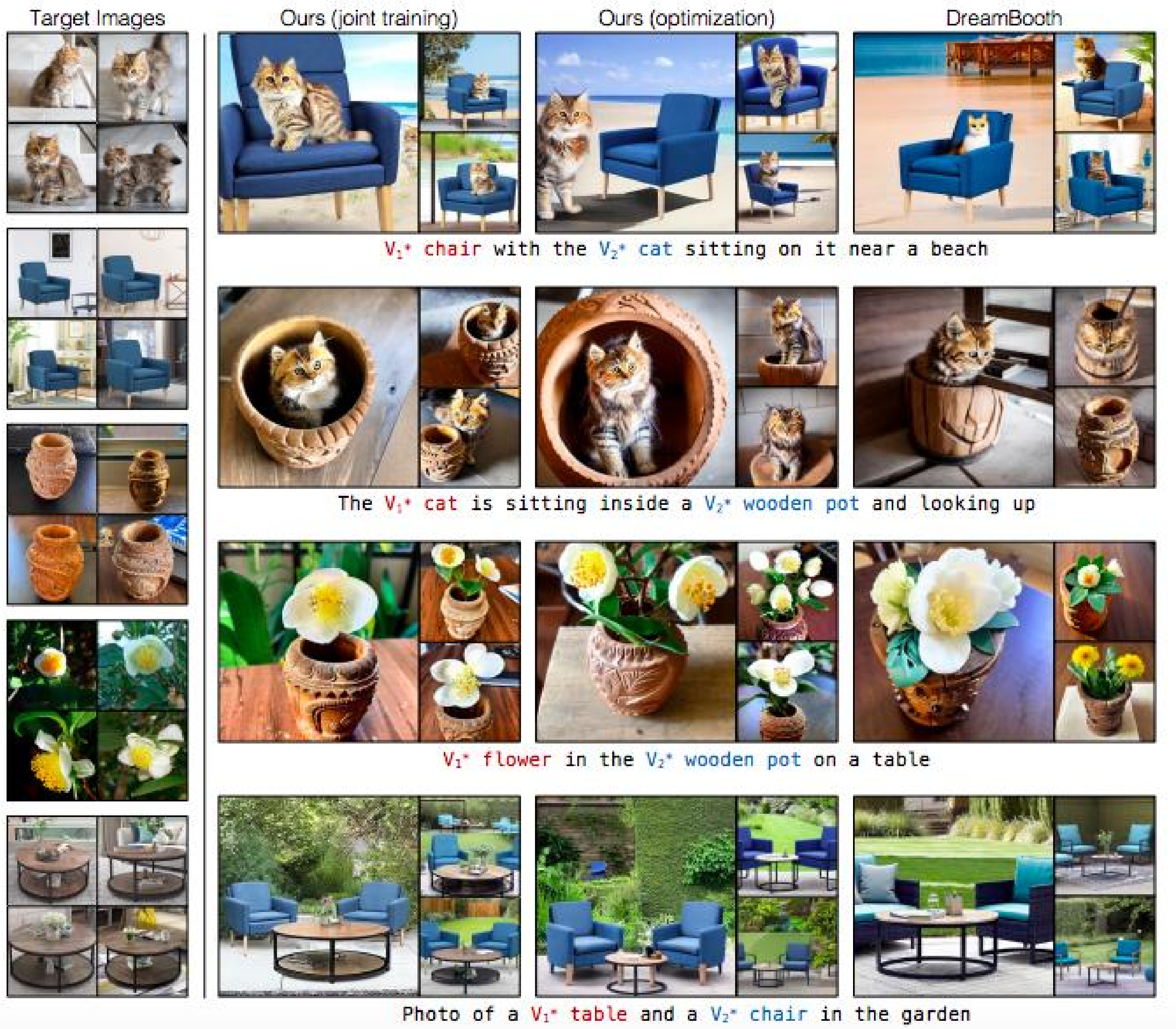
$$\hat{W} = \arg\min_{W} ||WC_{\text{reg}}^{\top} - W_0 C_{\text{reg}}^{\top}||_F$$
$$\text{s.t. } WC^{\top} = V, \text{ where } C = [\mathbf{c}_1 \cdots \mathbf{c}_N]^{\top}$$
$$\text{and } V = [W_1 \mathbf{c}_1^{\top} \cdots W_N \mathbf{c}_N^{\top}]^{\top}.$$

Solution

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

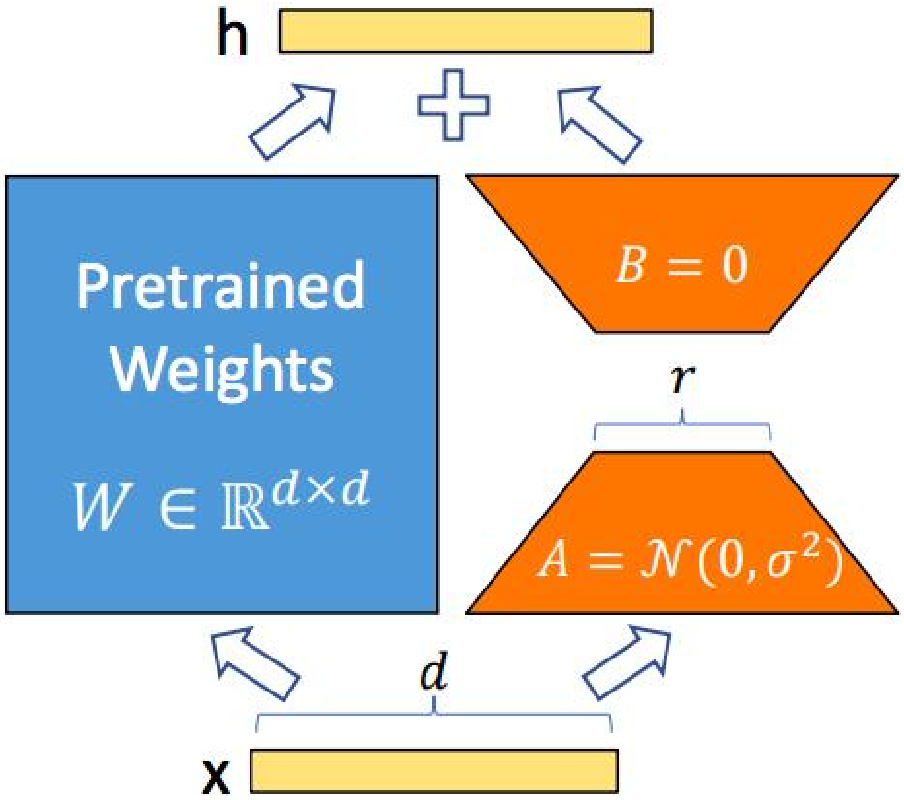
Background

- Comparison between two multi-concept composition methods



Background

- **LoRA: more efficient model fine-tuning**



Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

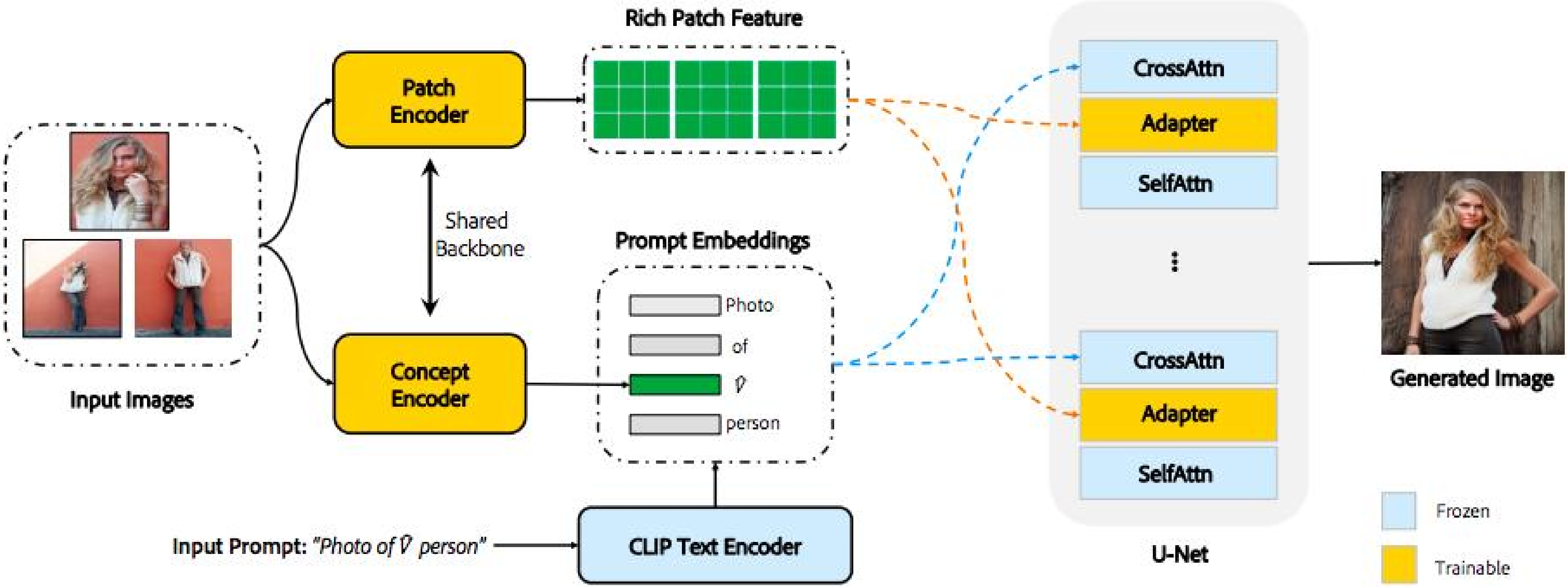
Background

Shi J, Xiong W, Lin Z, et al. Instantbooth: Personalized text-to-image generation without test-time finetuning[J]. arXiv preprint arXiv:2304.03411, 2023.

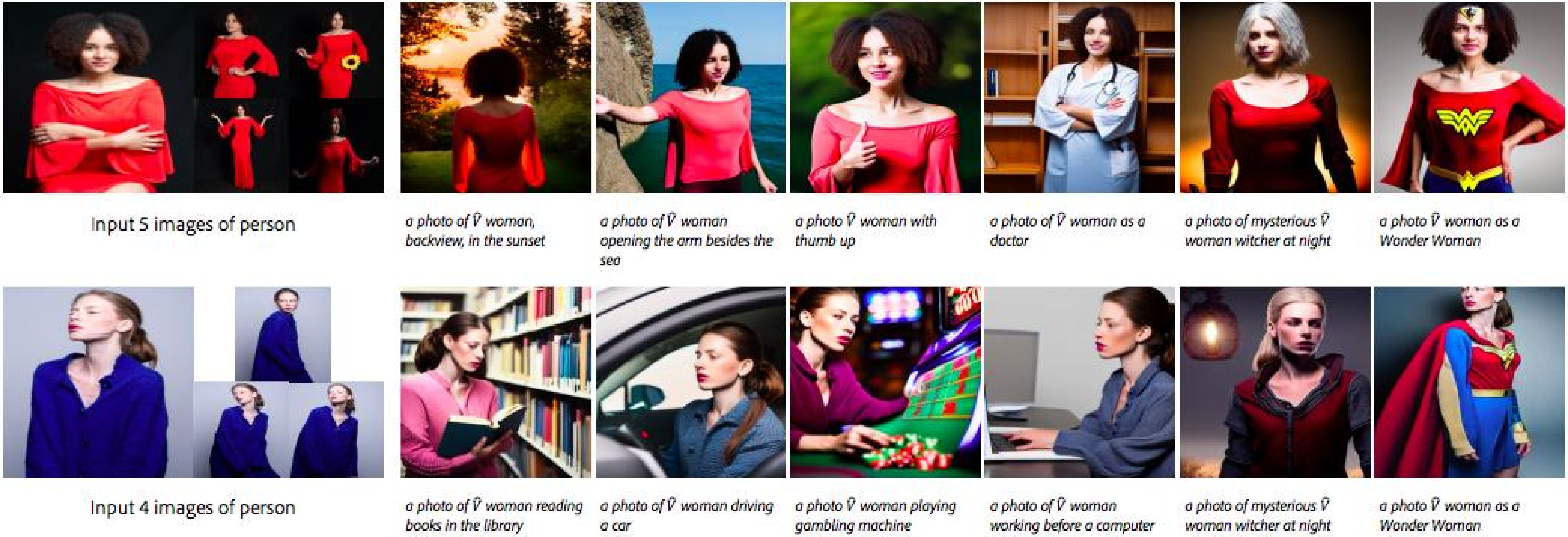
InstantBooth: eliminate the need for model fine-tuning

Background

Method overview



Background


Input 5 images of person
a photo of $\hat{V}$ woman, backview, in the sunset
a photo of $\hat{V}$ woman opening the arm besides the sea
a photo $\hat{V}$ woman with thumb up
a photo of $\hat{V}$ woman as a doctor
a photo of mysterious $\hat{V}$ woman witcher at night
a photo $\hat{V}$ woman as a Wonder Woman
Input 4 images of person
a photo of $\hat{V}$ woman reading books in the library
a photo of $\hat{V}$ woman driving a car
a photo $\hat{V}$ woman playing gambling machine
a photo of $\hat{V}$ woman working before a computer
a photo of mysterious $\hat{V}$ woman witcher at night
a photo $\hat{V}$ woman as a Wonder Woman

Background


Input 5 images of cat
a photo of $\hat{V}$ cat standing on the boat
a photo of $\hat{V}$ cat jumping on the floor
a photo $\hat{V}$ cat on the tree
a photo of $\hat{V}$ cat in a bucket
a watercolor painting of $\hat{V}$ cat
a photo $\hat{V}$ cat of on the piano
Input 5 images of cat
a photo of $\hat{V}$ cat wearing sunglasses on the beach
a photo of $\hat{V}$ cat in the swimming pool
a photo $\hat{V}$ cat of play with a ball
a photo of $\hat{V}$ cat in a bucket
a watercolor painting of $\hat{V}$ cat
a photo $\hat{V}$ cat of on the piano

Background

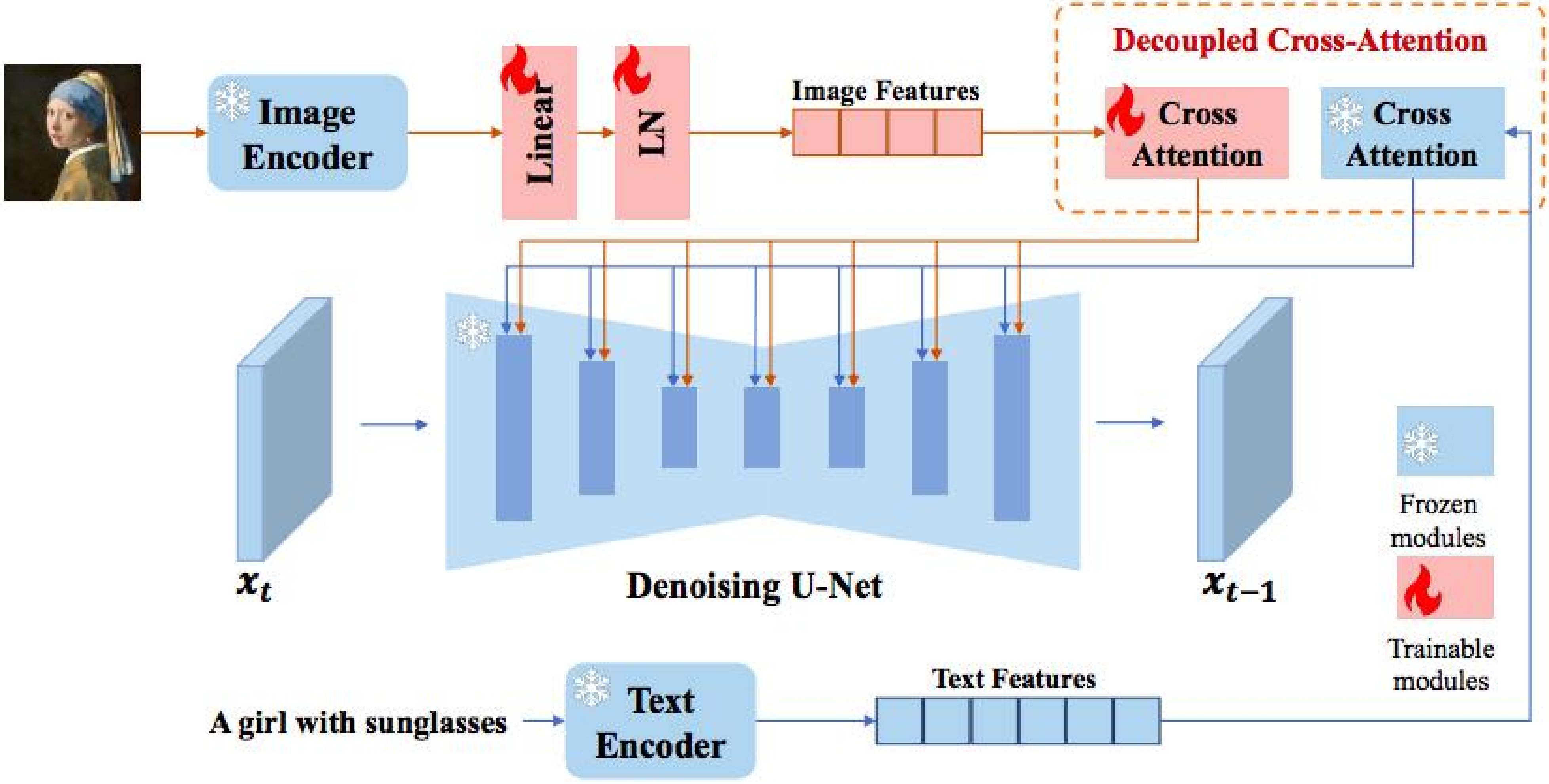
- **IP-Adapter: baseline of single image prompting**

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models

Hu Ye, Jun Zhang*, Sibo Liu, Xiao Han, Wei Yang
Tencent AI Lab
{huye, junejzhang, siboliu, haroldhan, willyang}@tencent.com

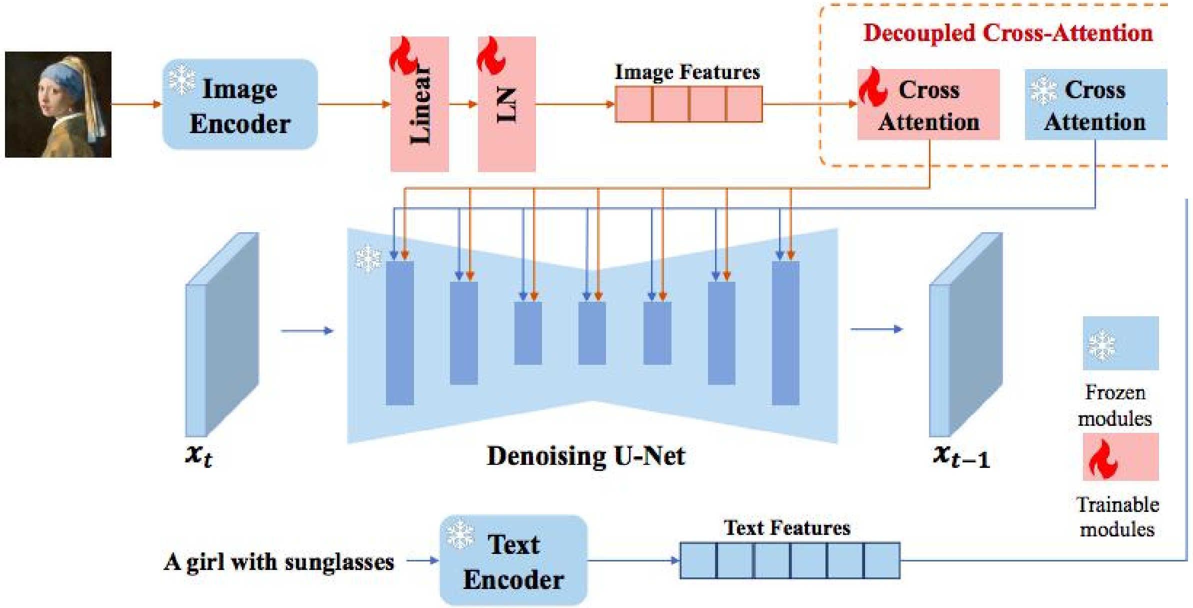
Background

Method overview



Background

Training objective



$$\mathbf{Z}^{new} = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} + \mathrm{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}'$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \mathbf{K} = \boldsymbol{c}_t\mathbf{W}_k, \mathbf{V} = \boldsymbol{c}_t\mathbf{W}_v, \mathbf{K}' = \boldsymbol{c}_i\mathbf{W}'_k, \mathbf{V}' = \boldsymbol{c}_i\mathbf{W}'_v$

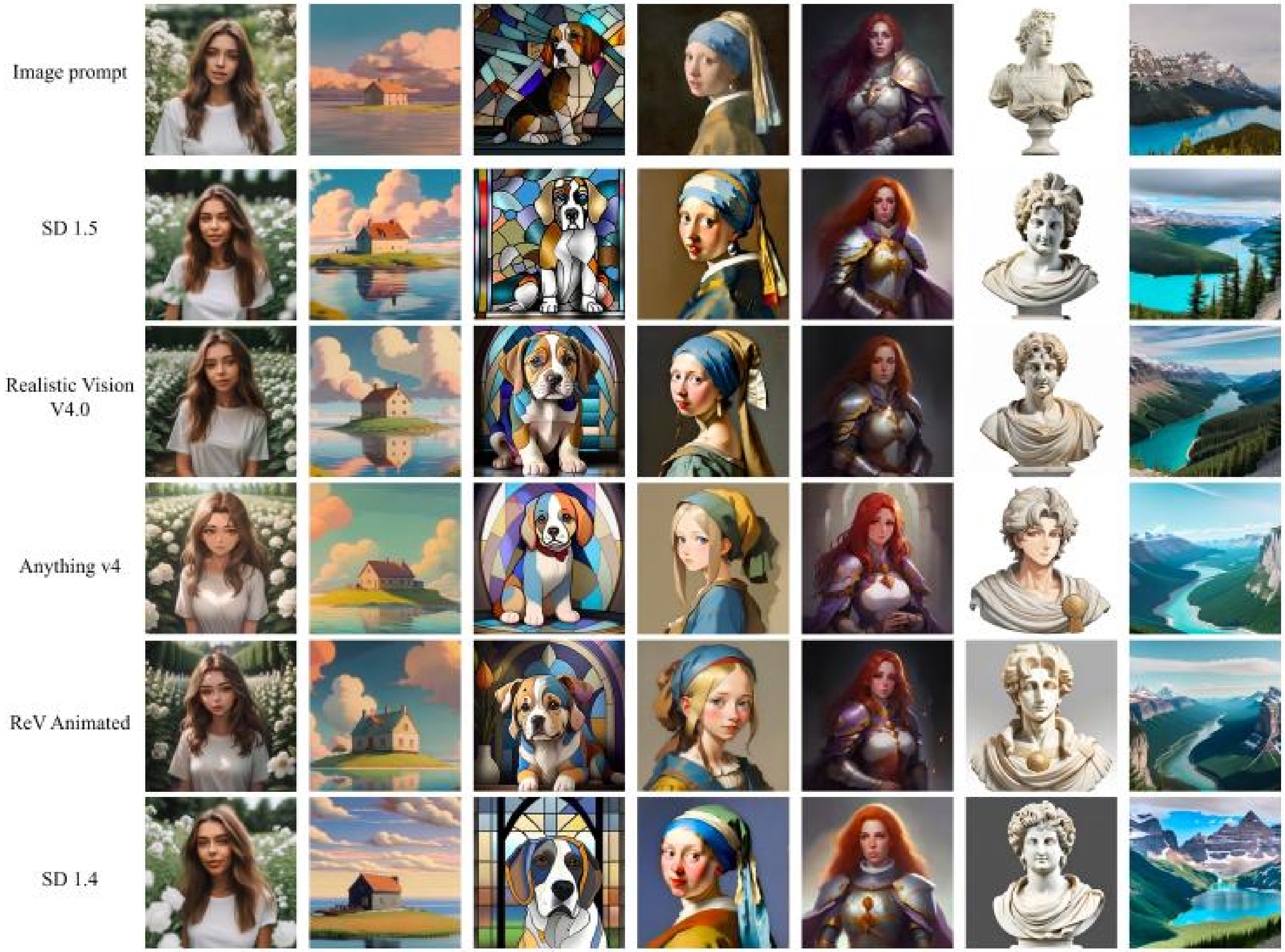
$$L_{\mathrm{simple}} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},\boldsymbol{c}_t,\boldsymbol{c}_i,t}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t)\|^2.$$

$$\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t) = w\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t) + (1-w)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$$

$$\mathbf{Z}^{new} = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \lambda \cdot \mathrm{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}')$$

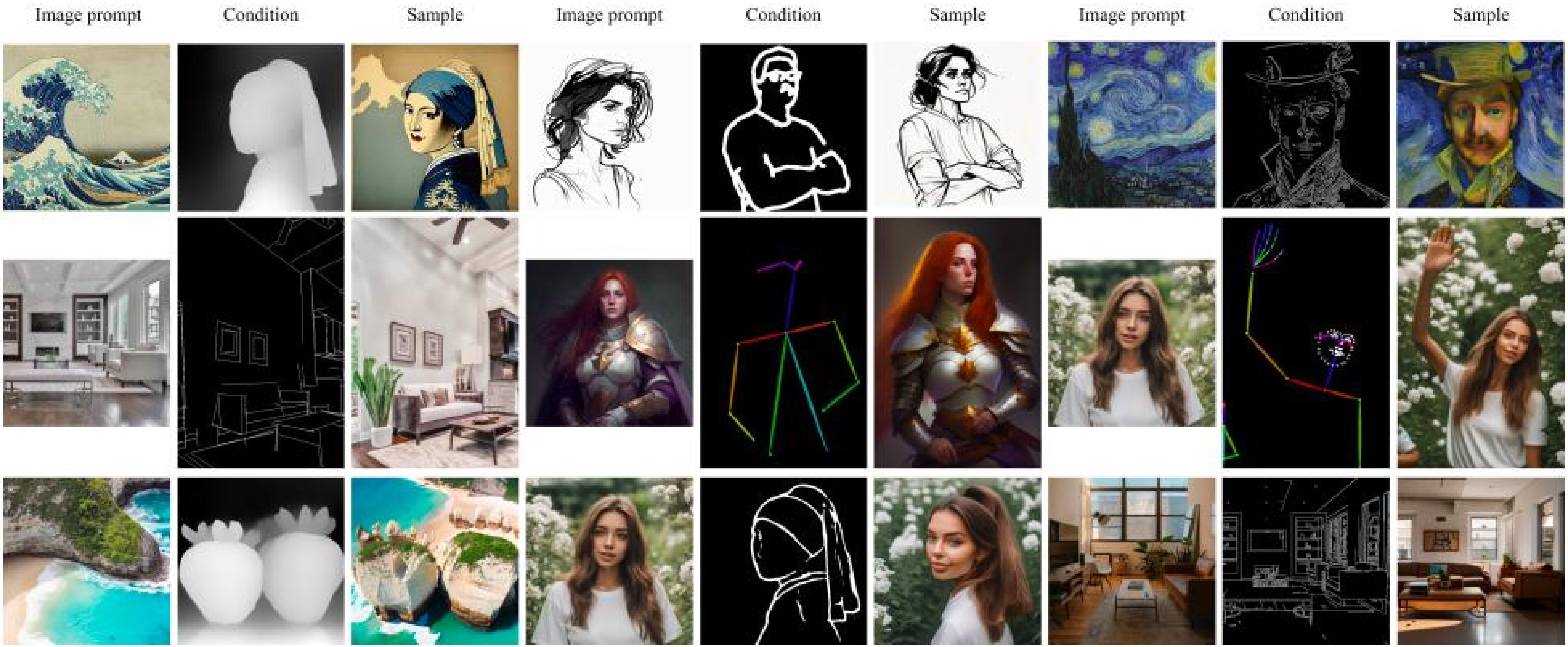
Background

- **Adaptation to different diffusion models when training only once**



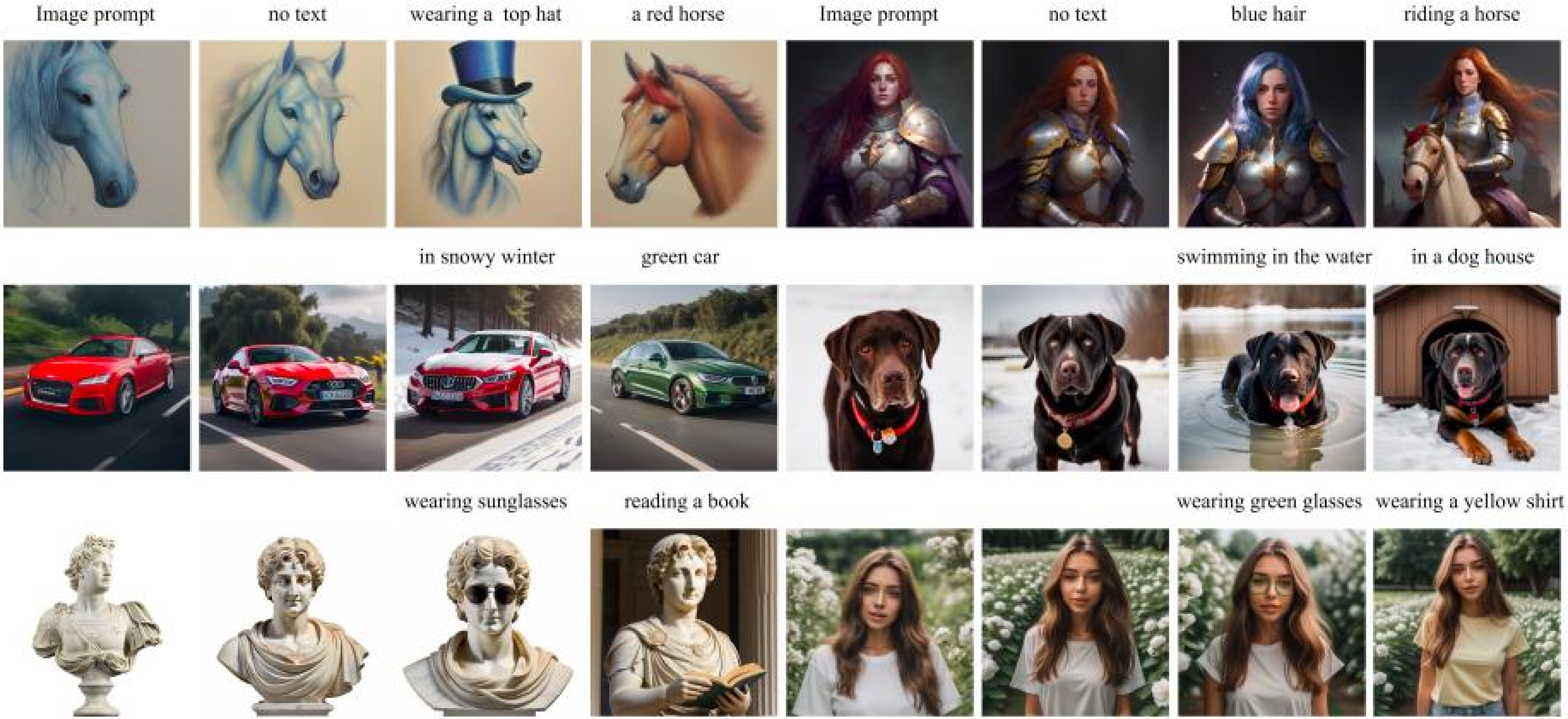
Background

Adaptation to ControlNet



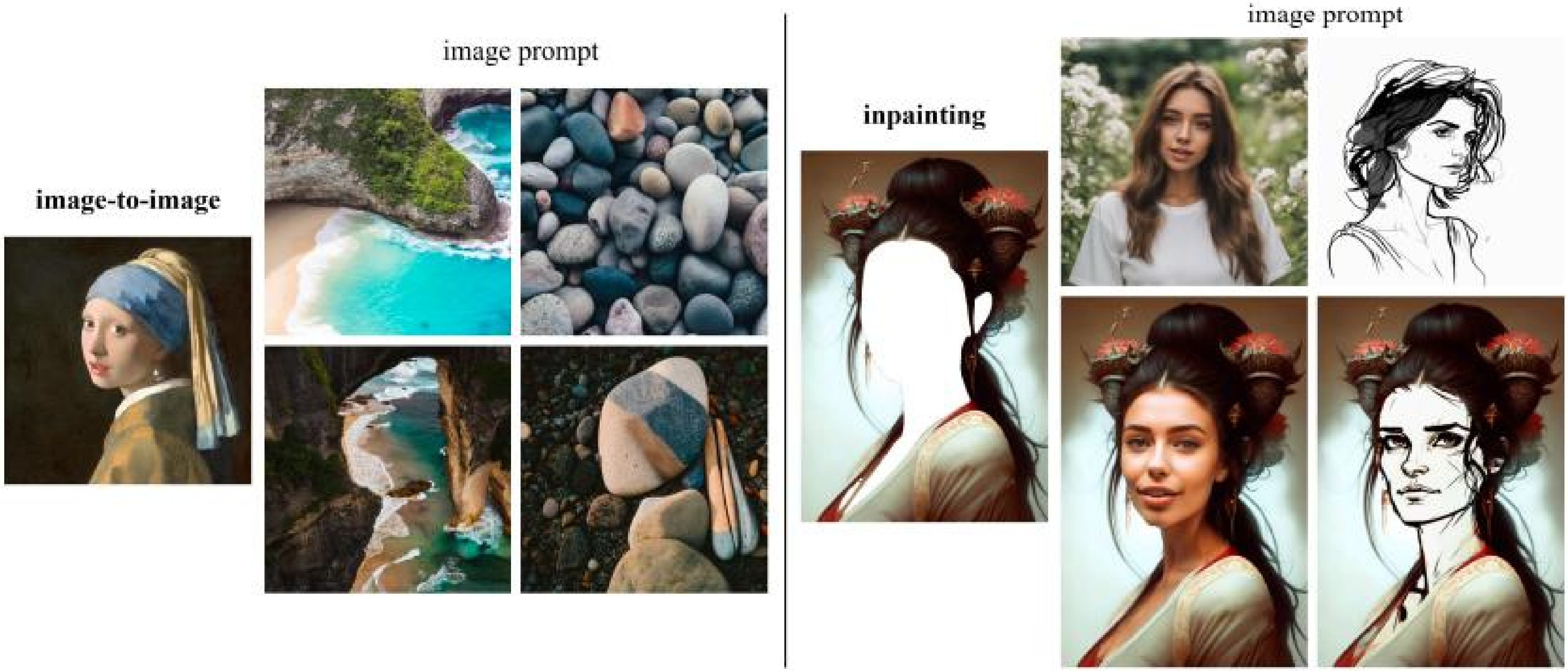
Background

Image prompt and text editing



Background

Application in I2I translation and inpainting



Background

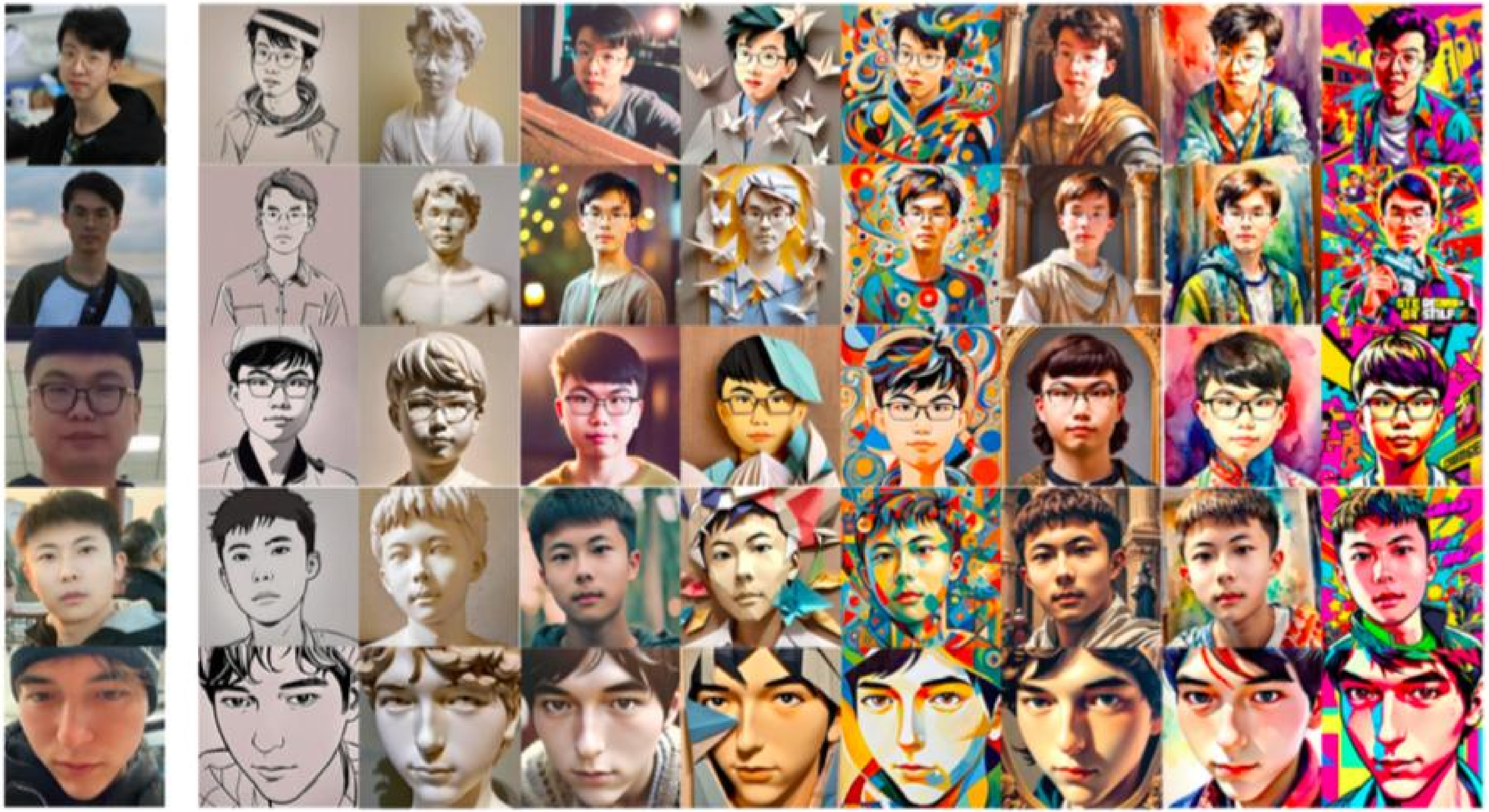
InstantID: Zero-shot Identity-Preserving Generation in Seconds

Qixun Wang[12], Xu Bai[12], Haofan Wang[12⋆], Zekui Qin[12], Anthony Chen[123], Huaxia Li[2], Xu Tang[2], and Yao Hu[2]

InstantX Team[1], Xiaohongshu Inc[2], Peking University[3]
{haofanwang.ai@gmail.com}
https://instantid.github.io

Background

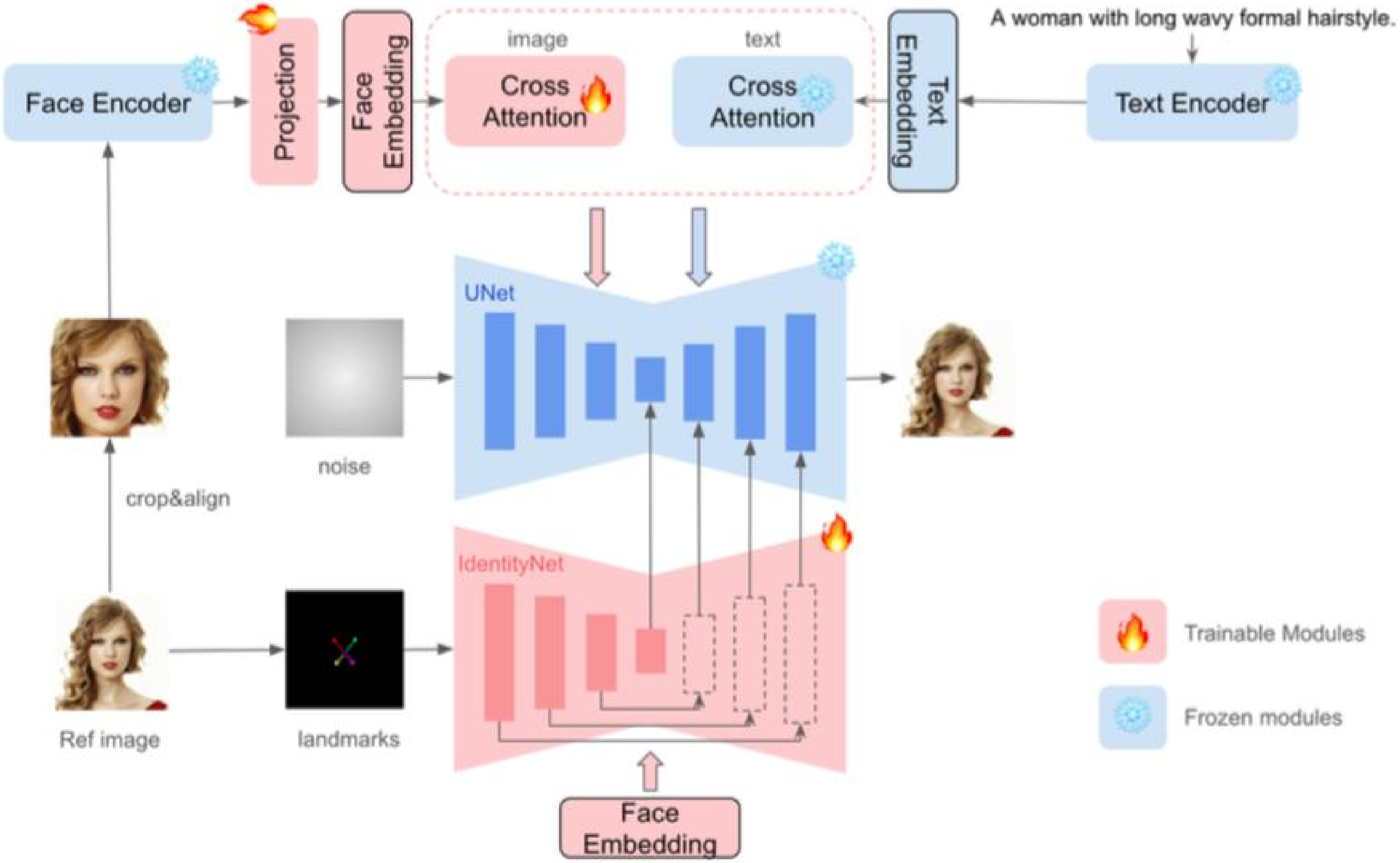
ID preserved T2I

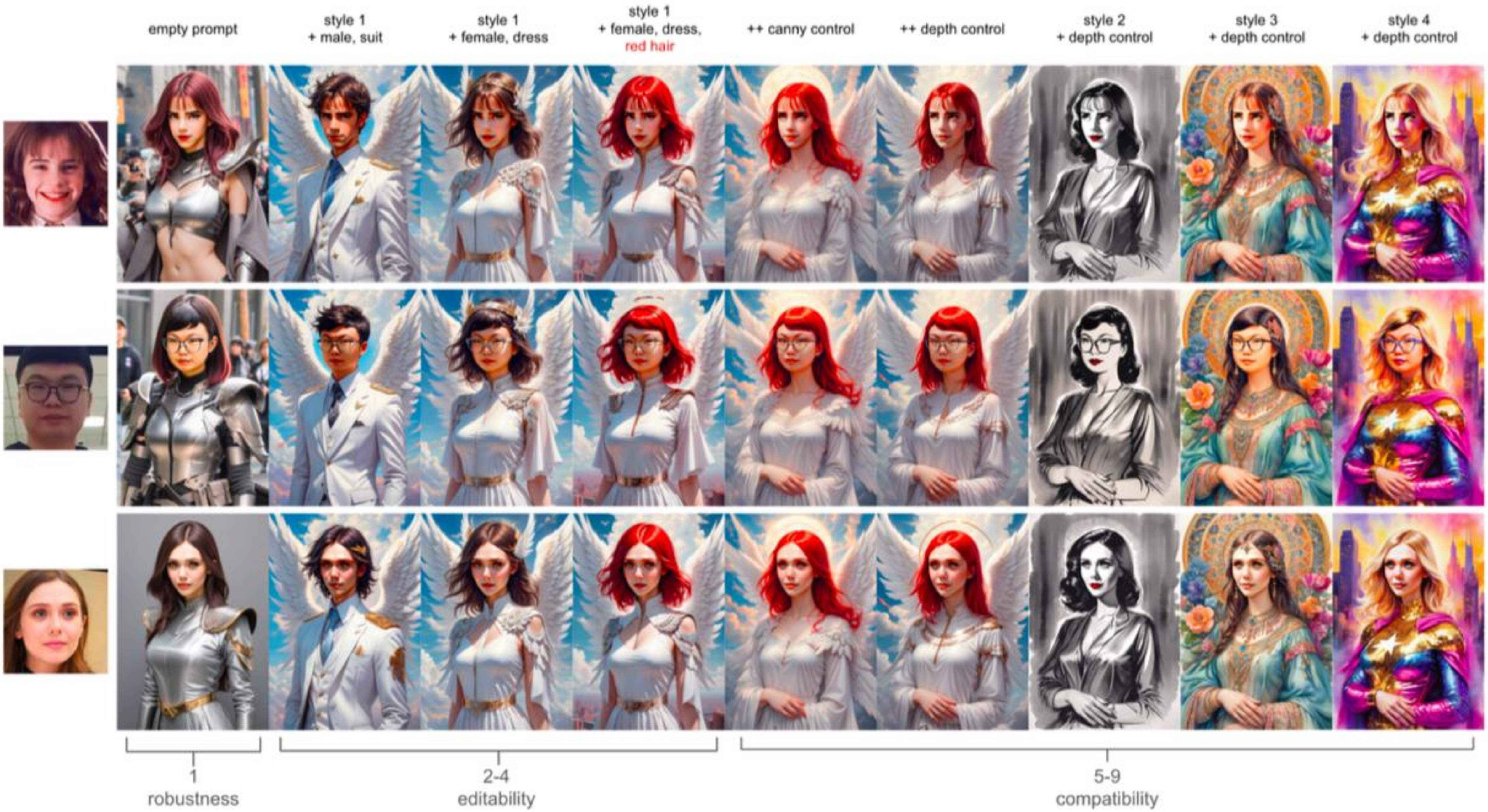


Background

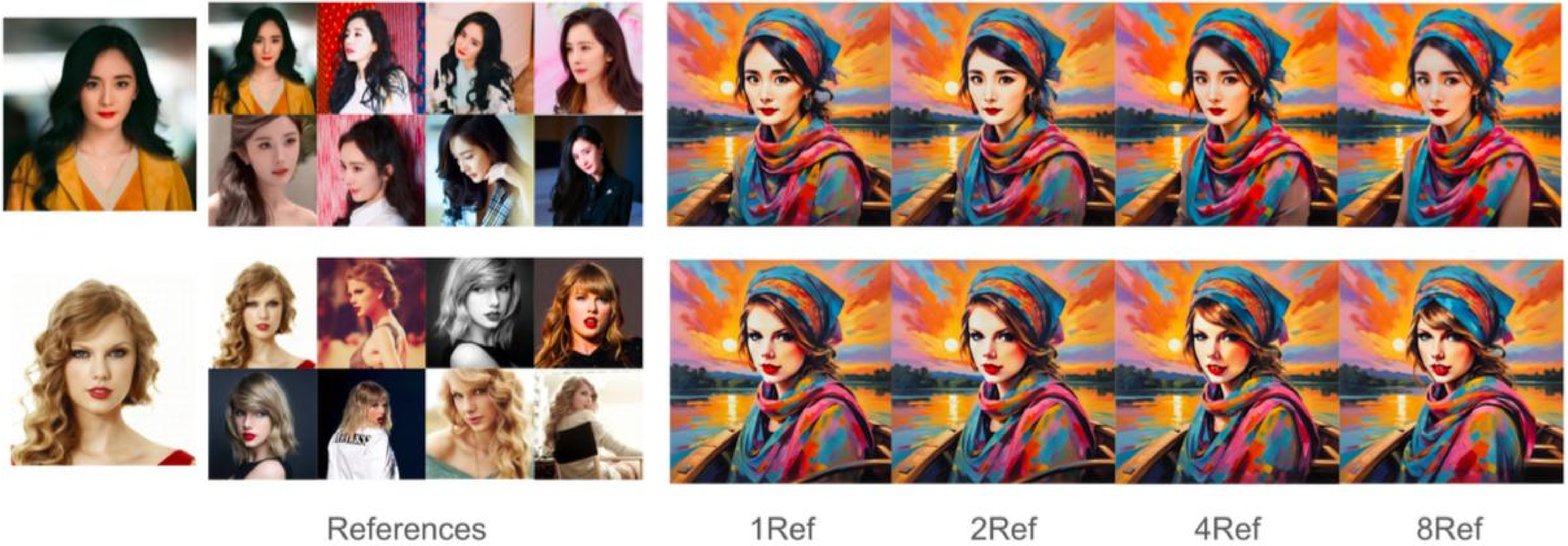
$$\mathcal{L} = \mathbb{E}_{z_t, t, C, C_i, \epsilon \sim \mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, C, C_i)||_2^2],$$

Method overview



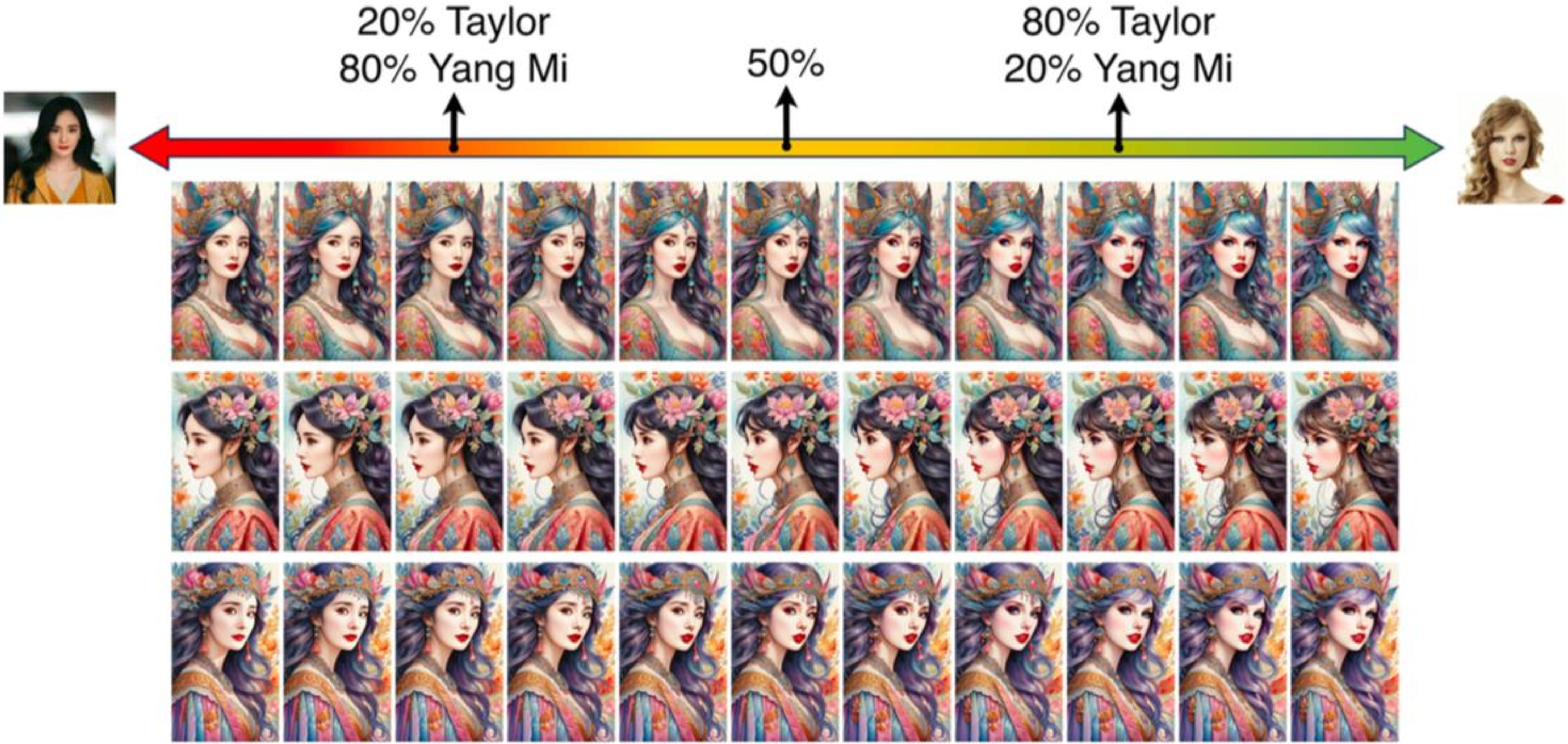

empty prompt
style 1
+ male, suit
style 1
+ female, dress
style 1
+ female, dress,
red hair
++ canny control
++ depth control
style 2
+ depth control
style 3
+ depth control
style 4
+ depth control
1
robustness
2-4
editability
5-9
compatibility

Background



Background

Pose control effect


20% Taylor
80% Yang Mi
50%
80% Taylor
20% Yang Mi

Background

CapHuman: Capture Your Moments in Parallel Universes

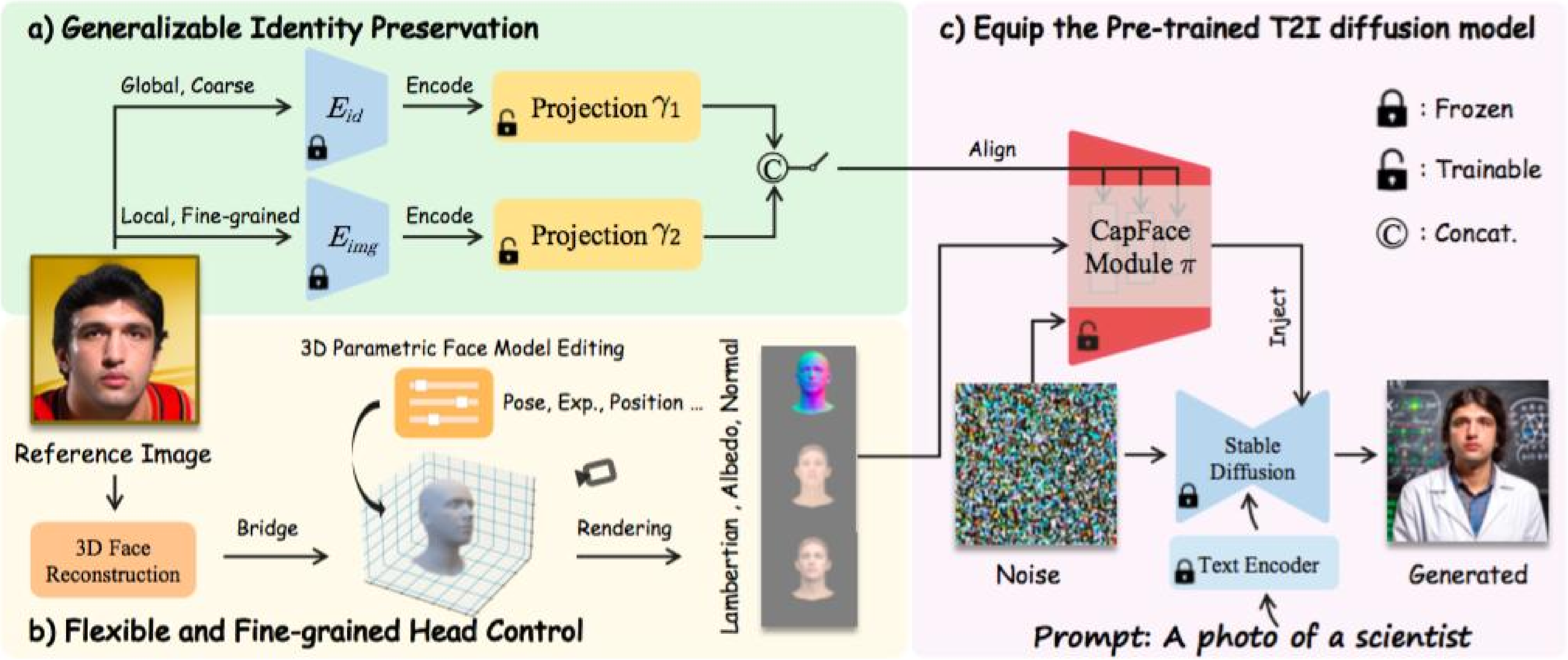
Chao Liang[1] Fan Ma[1] Linchao Zhu[1] Yingying Deng[2] Yi Yang[1†]
[1]ReLER, CCAI, Zhejiang University, Zhejiang, China [2]Huawei Technologies Ltd., China
[†] Corresponding author
{cs.chaoliang, zhulinchao, yangyics}@zju.edu.cn, flower.fan@foxmail.com, dyy15@outlook.com
https://caphuman.github.io

Background


Reference Image
your first life
your second life
your third life
3D facial prior
... a pop singer, sing, play the guitar, piano, take part in the show
... a scientist, work with Hawking, Hinton, present in a conference
... an astronaut, travel over the universe, collaborate with Obama

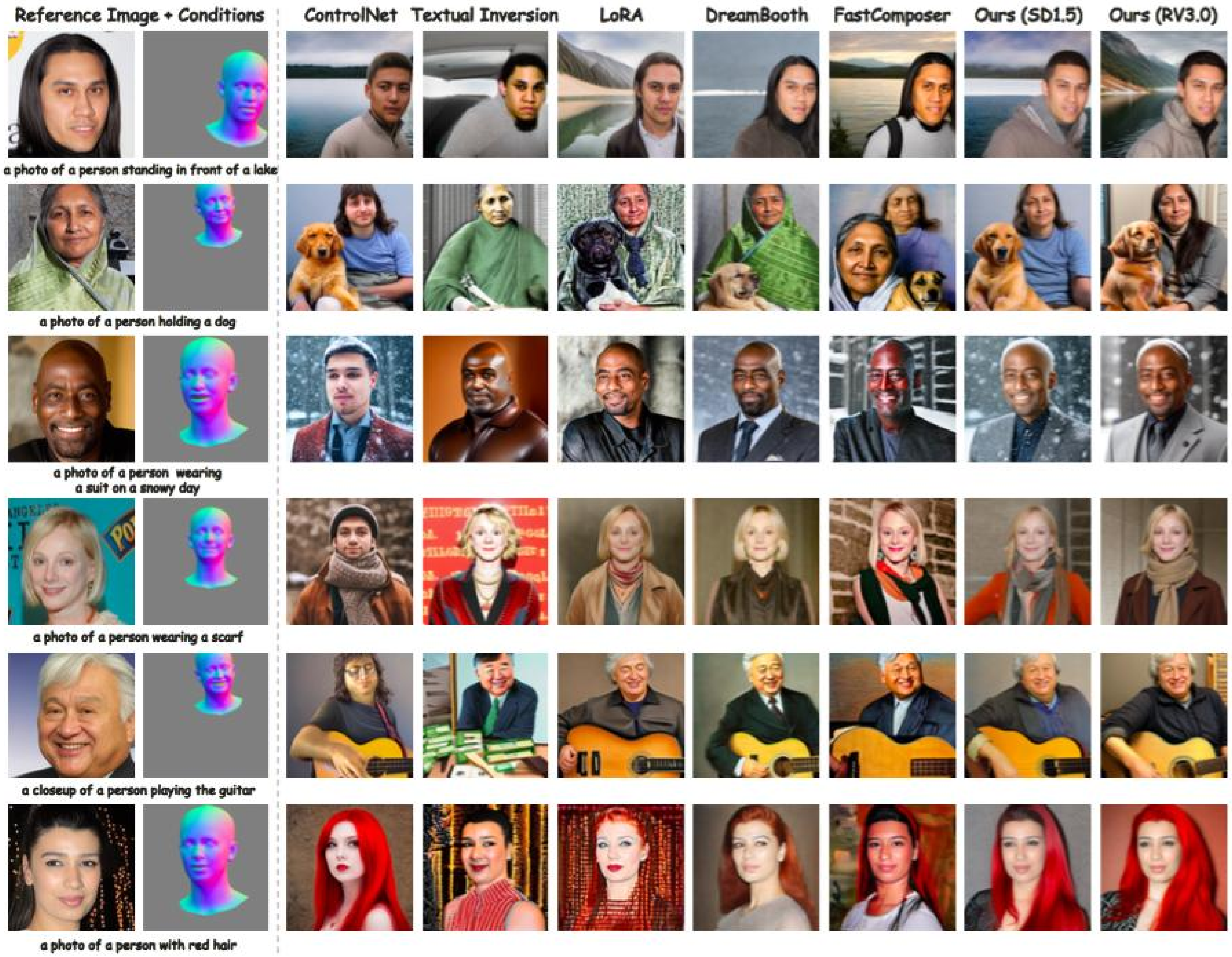
Background

Method overview



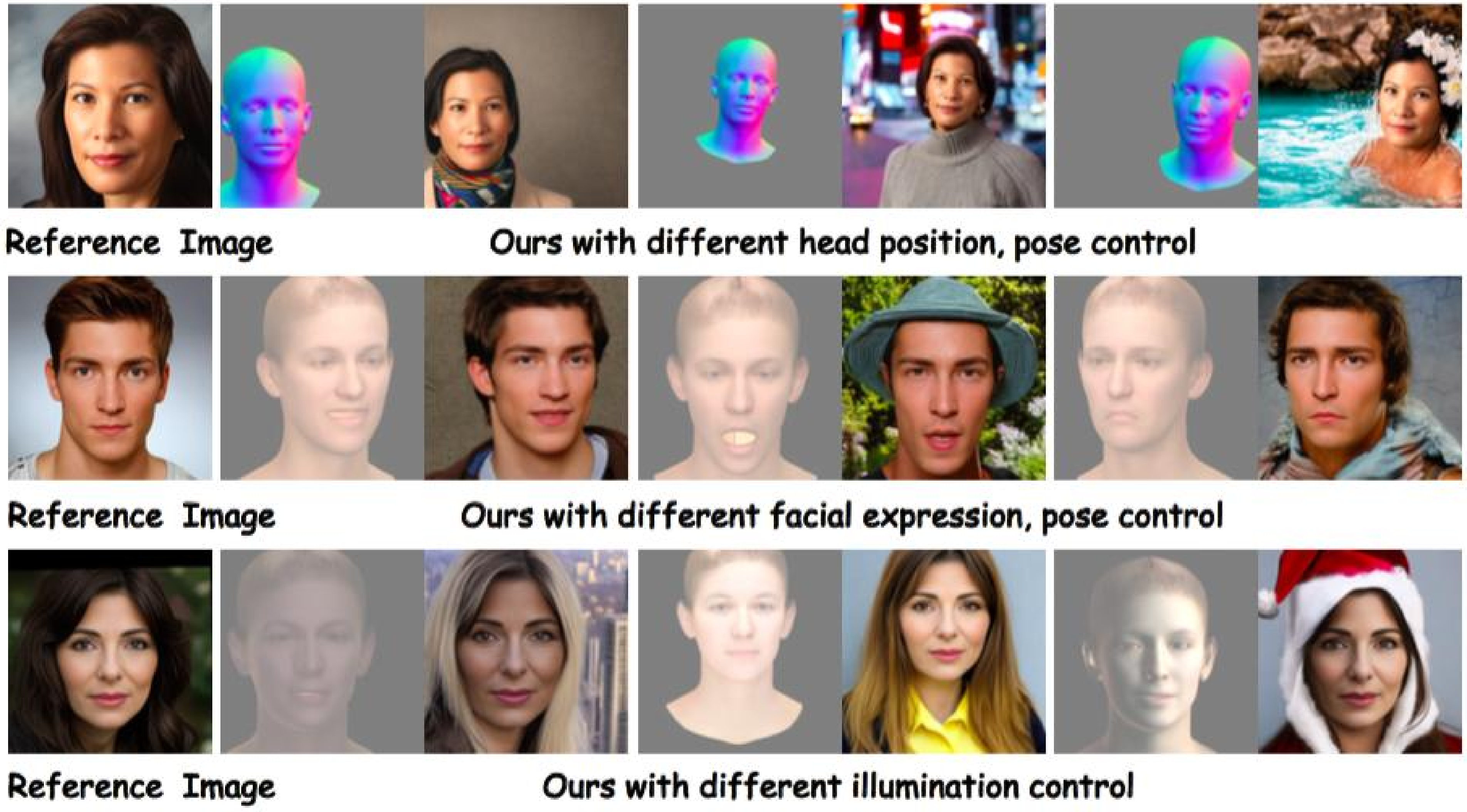
Background

Qualitative results



Background

More control effects of 3dMM



Background

InstantFamily: Masked Attention for Zero-shot Multi-ID Image Generation

Chanran Kim
SK Telecom
Seoul, Republic of Korea
chanrankim@sk.com

Jeongin Lee
SK Telecom
Seoul, Republic of Korea
jeonginlee@sk.com

Shichang Joung
SK Telecom
Seoul, Republic of Korea
shichang.joung@sk.com

Bongmo Kim
SK Telecom
Seoul, Republic of Korea
bongmo.kim@sk.com

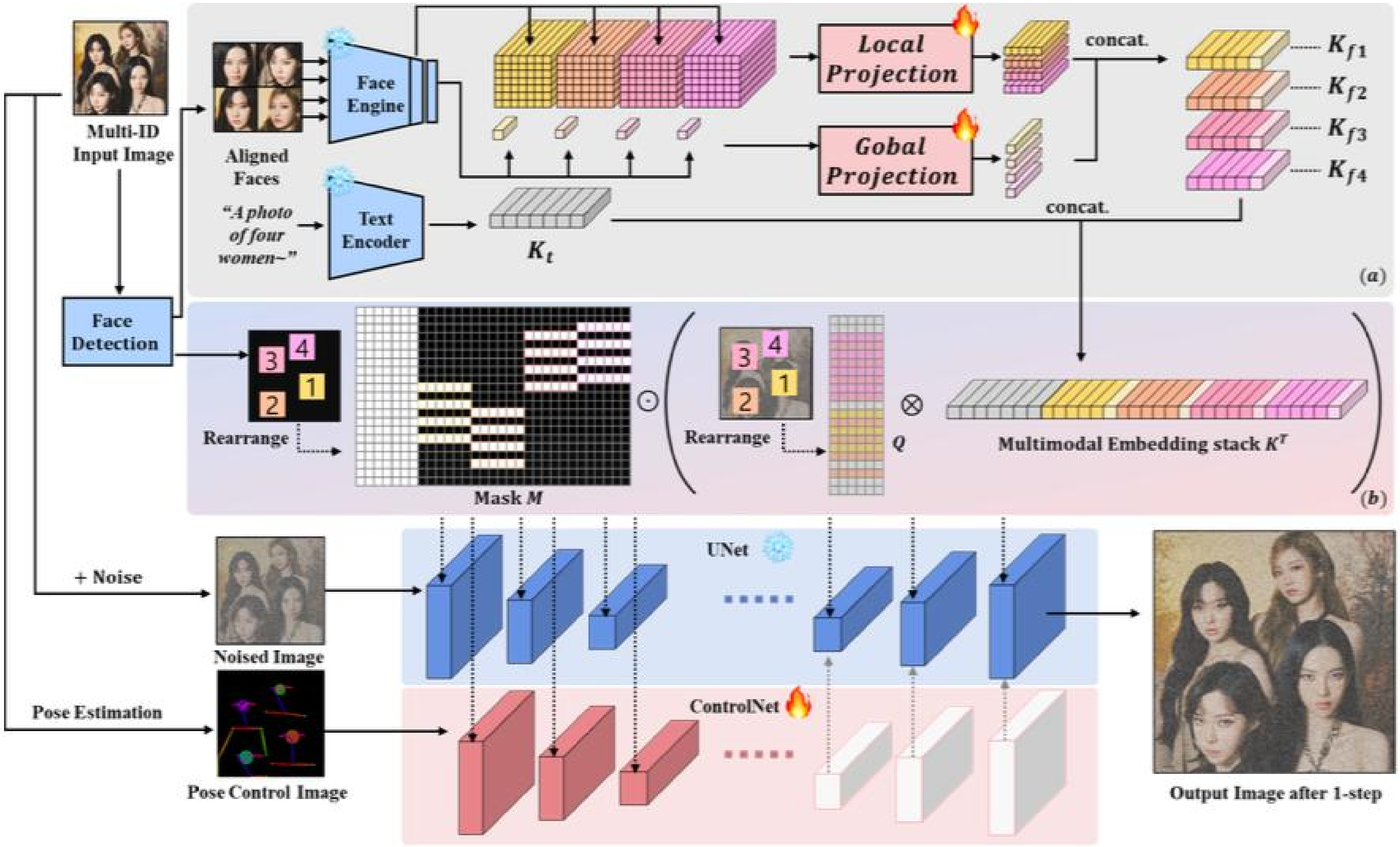
Yeul-Min Baek
SK Telecom
Seoul, Republic of Korea
ym.baek@sk.com

Background


Multi-ID Input
Text Input
"A photo
of seven men
in mars ~"
Pose Input

Background

Method overview




Pose Control Image
Cherry Blossom
Jungle
Studio


1 ID
2 IDs
3 IDs
4 IDs
4 IDs
Street
School
Mars

Outline

1 Background

2 Author

3 Method

4 Experiments

Author

MoMA: Multimodal LLM Adapter for Fast Personalized Image Generation

Kunpeng Song[1,2], Yizhe zhu[1], Bingchen Liu[1], Qing Yan[1], Ahmed Elgammal[2], and Xiao Yang[1]

[1] ByteDance
[2] Rutgers University

Author

KUNPENG SONG

Rutgers University
Verified email at cs.rutgers.edu

Computer Vision Deep Learning Machine Learning AIGC

First author: PhD student in Rutgers University, major in computer vision, AIGC

Author

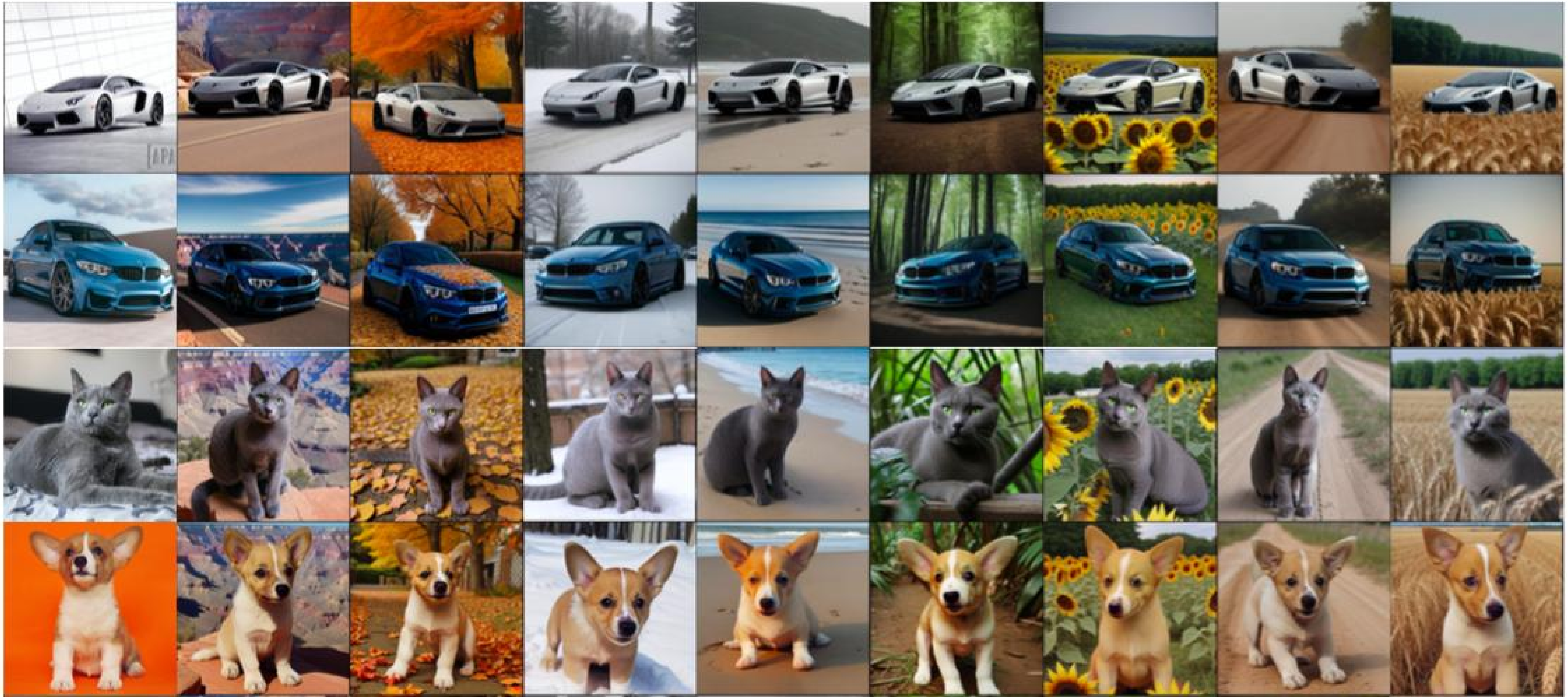
艾哈迈德·埃尔加马尔 (Ahmed Elgammal)

Dr. Ahmed Elgammal is a professor at the Department of Computer Science at Rutgers University. He is the founder and director of the Art and Artificial Intelligence Laboratory at Rutgers, which focuses on data science in the domain of digital humanities.

Dr. Elgammal received his M.Sc. and Ph.D. degrees in computer science from the University of Maryland, College Park, in 2000 and 2002, respectively.

Method

Results preview

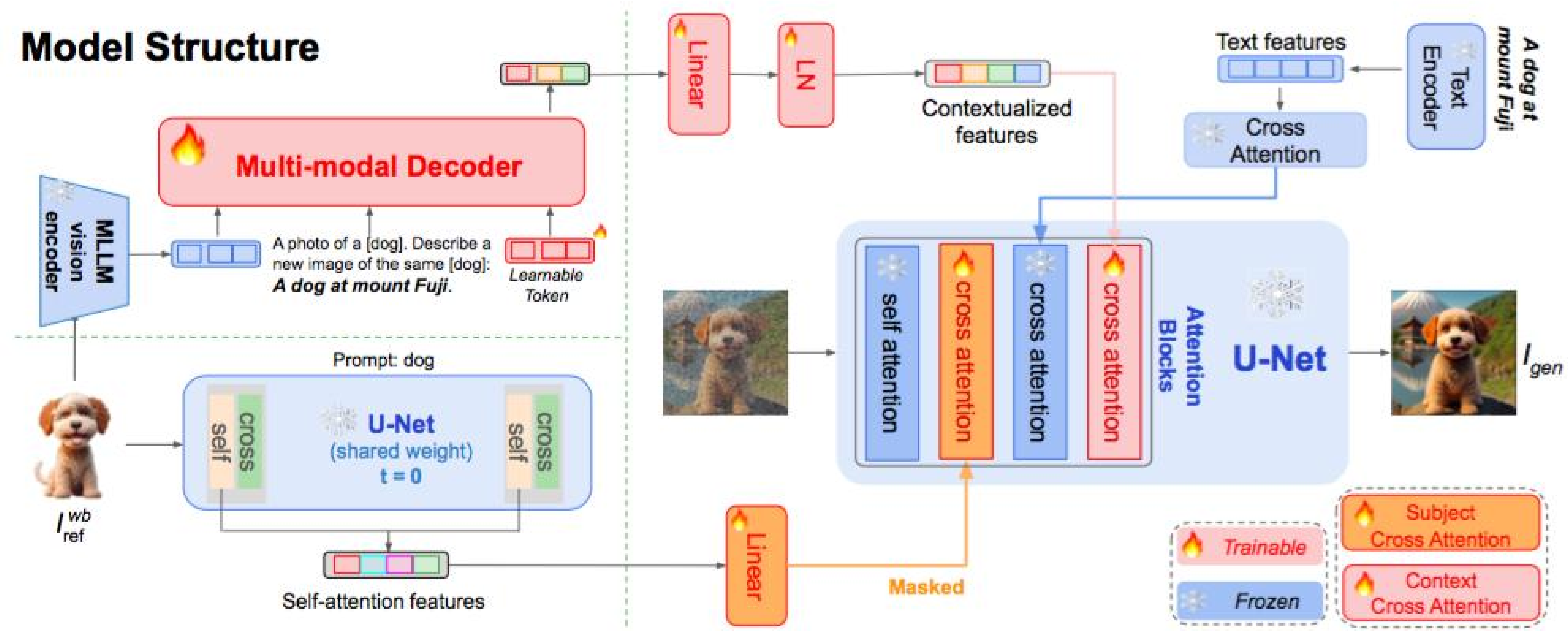


Method

Contributions

- Mask based self-supervised multi-modal generative learning
- Introduction of MLLM for better feature learning
- Disentangled cross-attention and self-attention
- Iterative self-attention masking

Method



Model Structure
Multi-modal Decoder
MLLM vision encoder
A photo of a [dog]. Describe a new image of the same [dog]: A dog at mount Fuji.
Learnable Token
Prompt: dog
U-Net (shared weight) t = 0
self
cross
Self-attention features
Linear
LN
Contextualized features
Text features
Text Encoder
A dog at mount Fuji
Cross Attention
self attention
cross attention
cross attention
cross attention
Attention Blocks
U-Net
Masked
Trainable
Frozen
Subject Cross Attention
Context Cross Attention

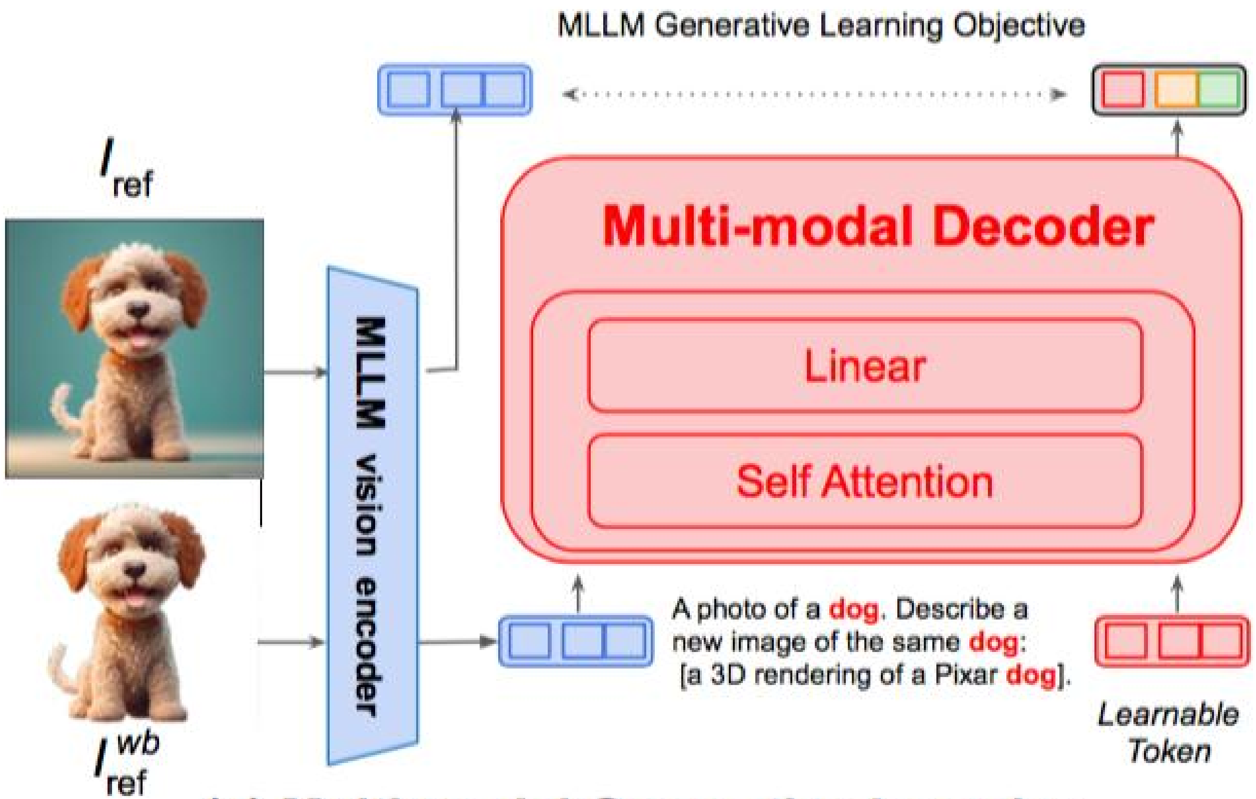
Method

Stage 1:

multi-modal generative learning

LLaVA

19.Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)

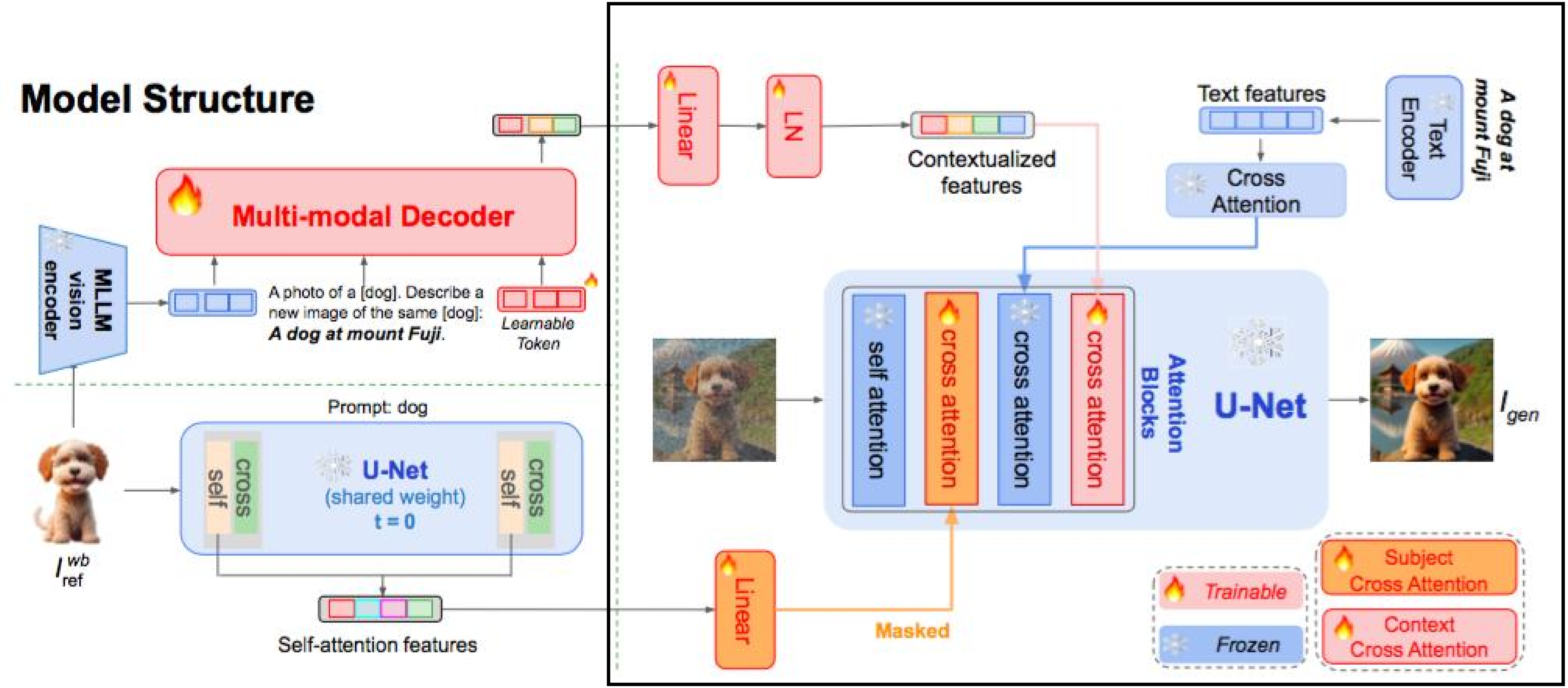


(a) Multi-modal Generative Learning

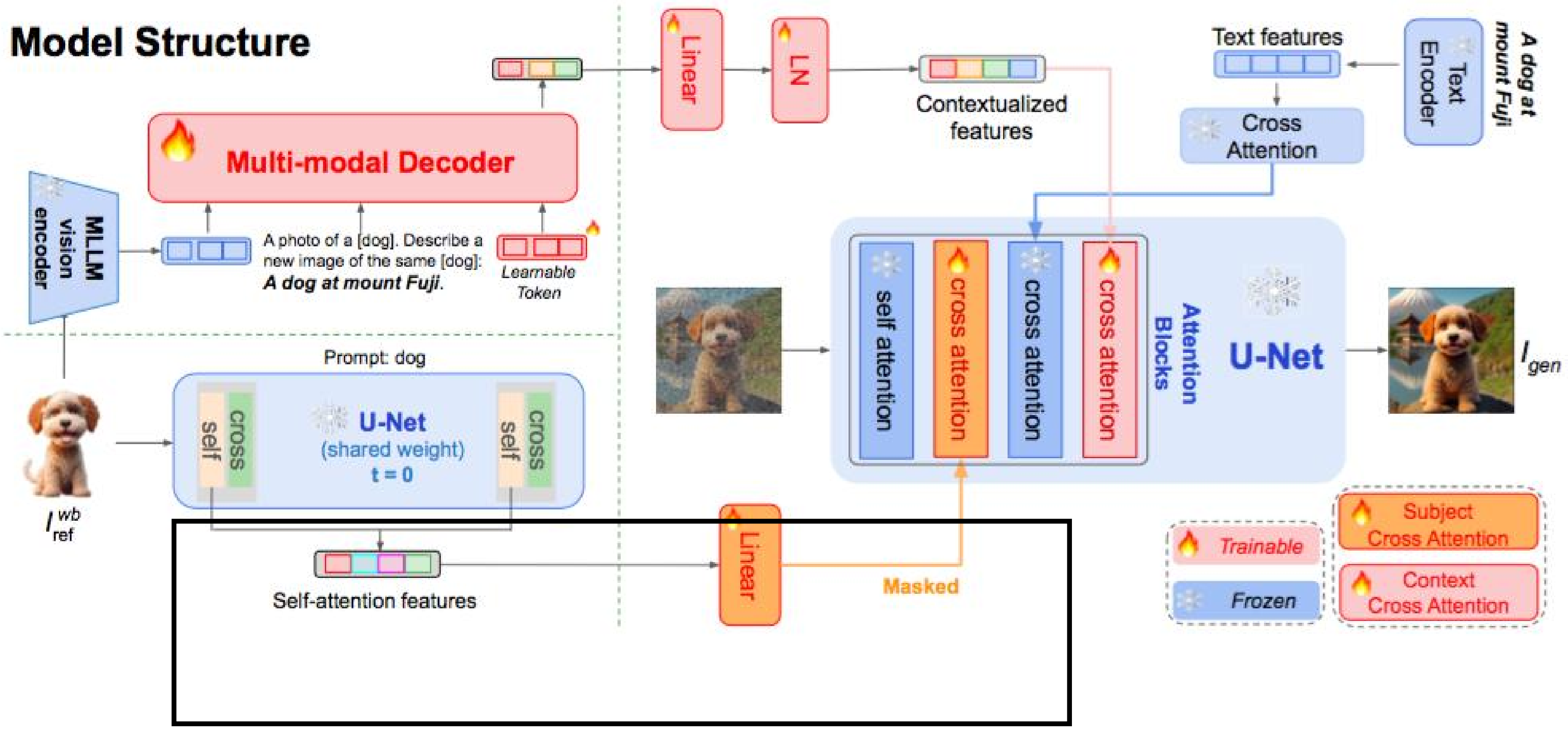
$$\mathcal{L}_{\text{MLLM}} = \left\| \text{MLLM}\left(\text{CLIP}\left(I^{wb}_{ref}\right), \text{P}_{ref}, \text{Token}\right) - \text{CLIP}\left(I_{ref}\right) \right\|_2^2$$

Method

Stage 2



Method

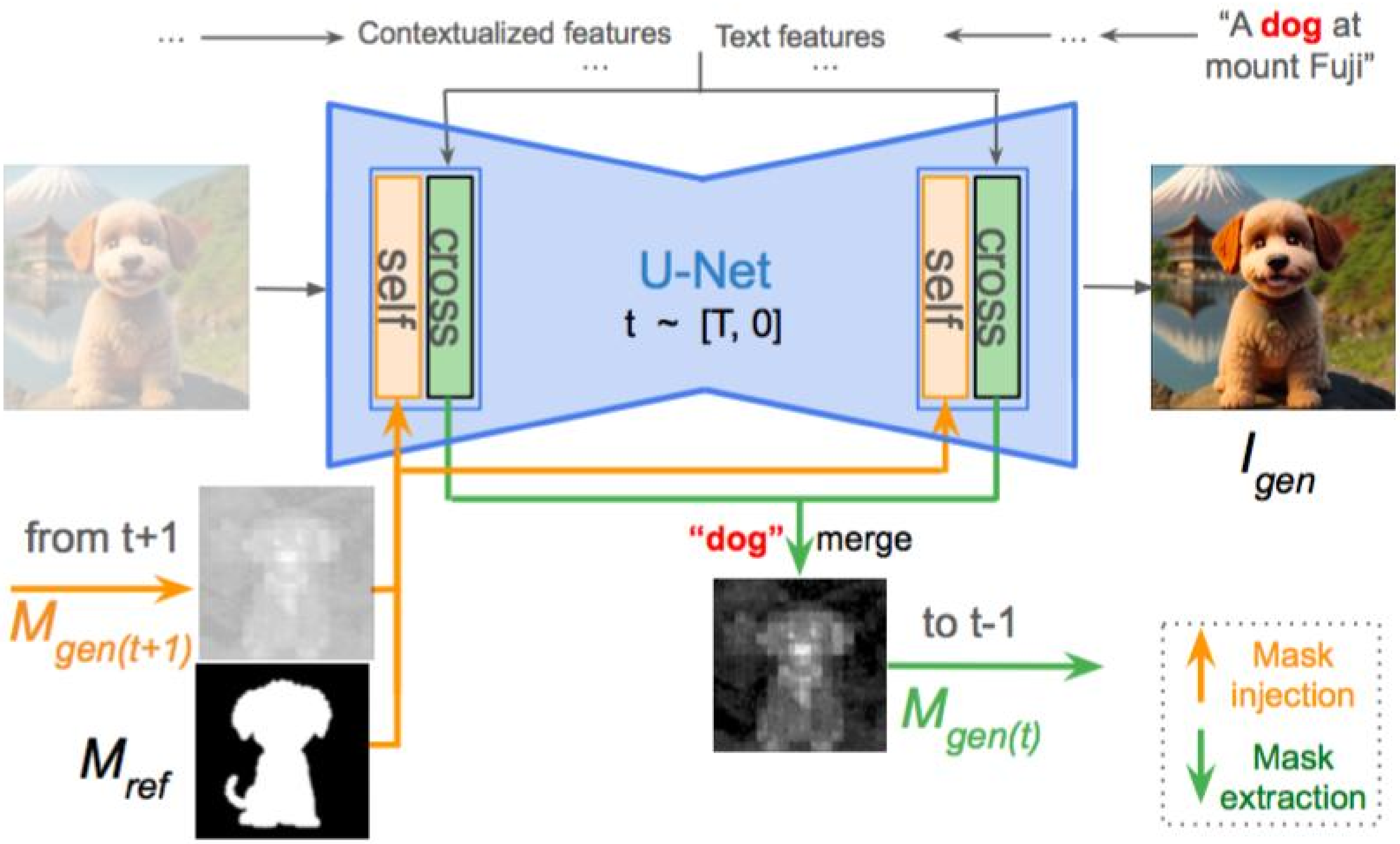


$$Z_{new} = Attn(Q, K, V) + \lambda \cdot Attn(Q, K', V', M_{ref}) \cdot M_{gen} \cdot \beta$$

Method

Inference:

Approximate generated mask with the cross attention map



(b) Iterative Self Attention Masking

Outline

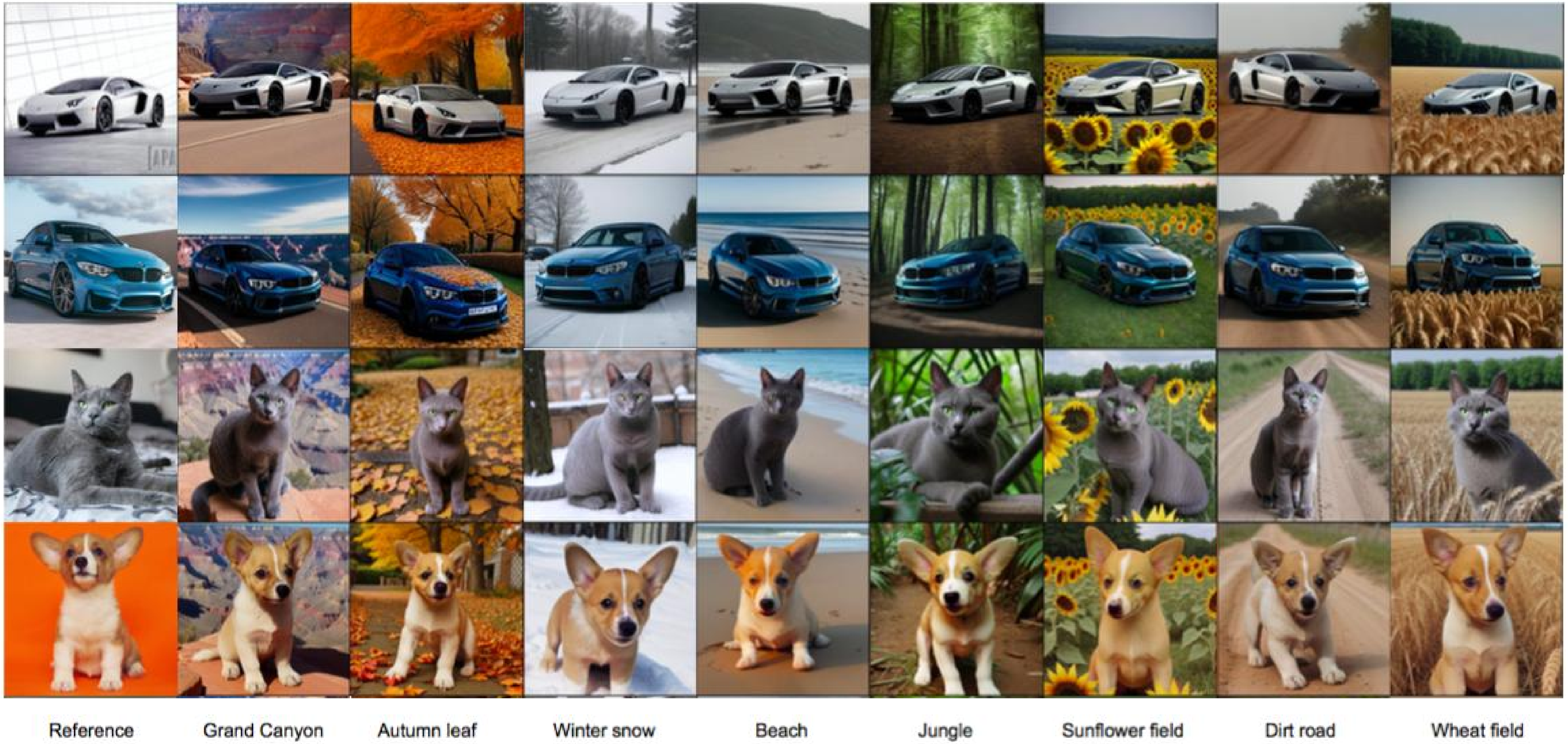
1 Background

2 Author

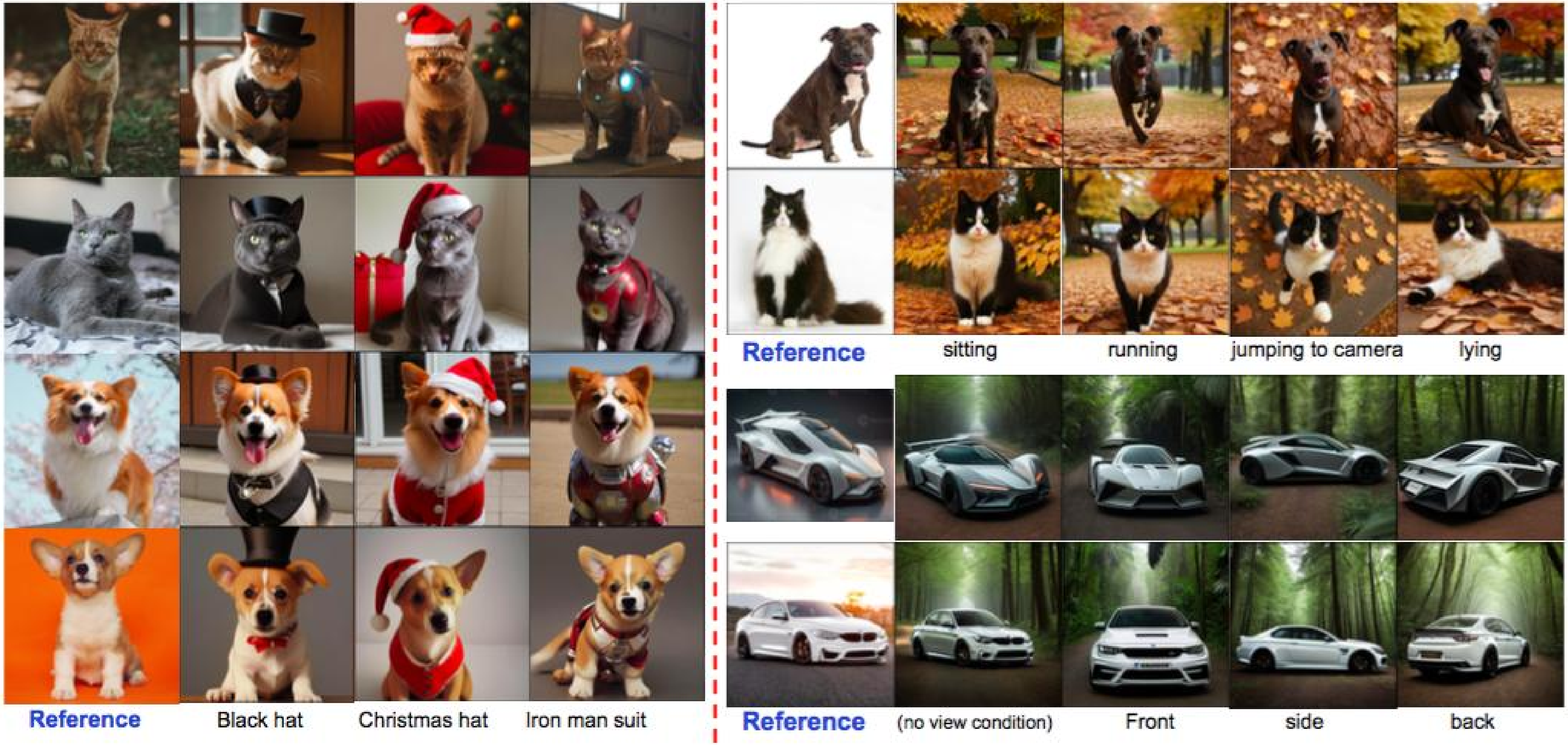
3 Method

4 Experiments

Experiments



Experiments



Experiments

Adapt to pre-trained community models

Thanks!