STRUCT Group Paper Reading

DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations

Tianhao Qi[1*], Shancheng Fang[1], Yanze Wu[2†], Hongtao Xie[1✉], Jiawei Liu[2],
Lang Chen[2], Qian He[2], Yongdong Zhang[1]
[1]University of Science and Technology of China, [2]ByteDance,
[*]Works done during the internship at ByteDance, [†]Project lead, [✉]Corresponding author,

PRESENTER: GUO TANG

2024/5/19

Outline

1 Background

2 Author

3 Method

4 Experiments

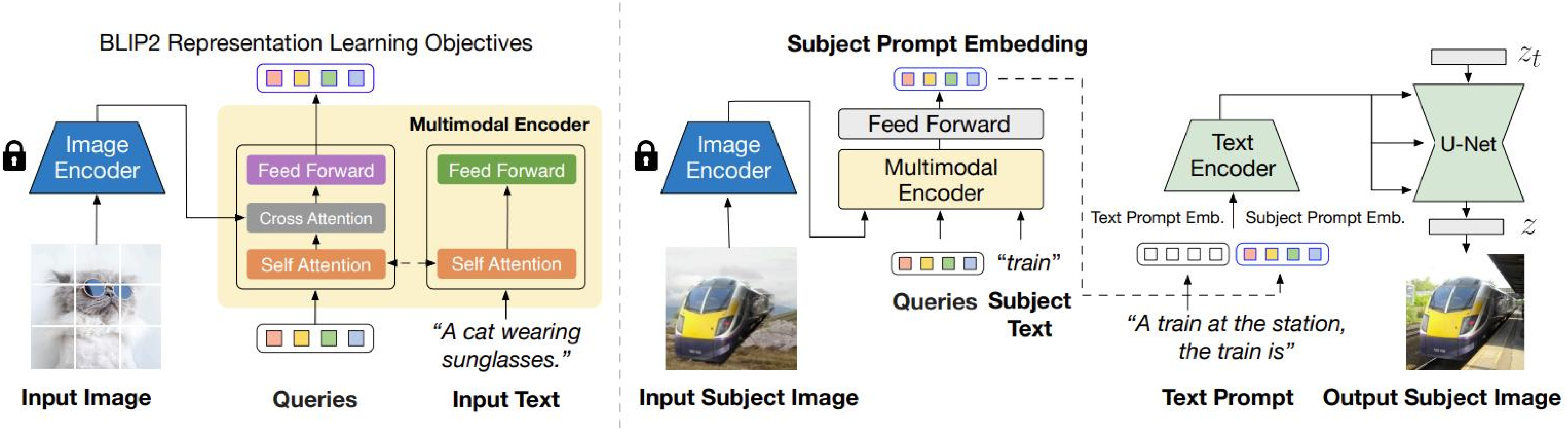
Background

BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing

Dongxu Li†, **Junnan Li**†, **Steven C.H. Hoi**†
Salesforce AI Research
†Corresponding authors: {li.d,junnan.li,shoi}@salesforce.com
https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion

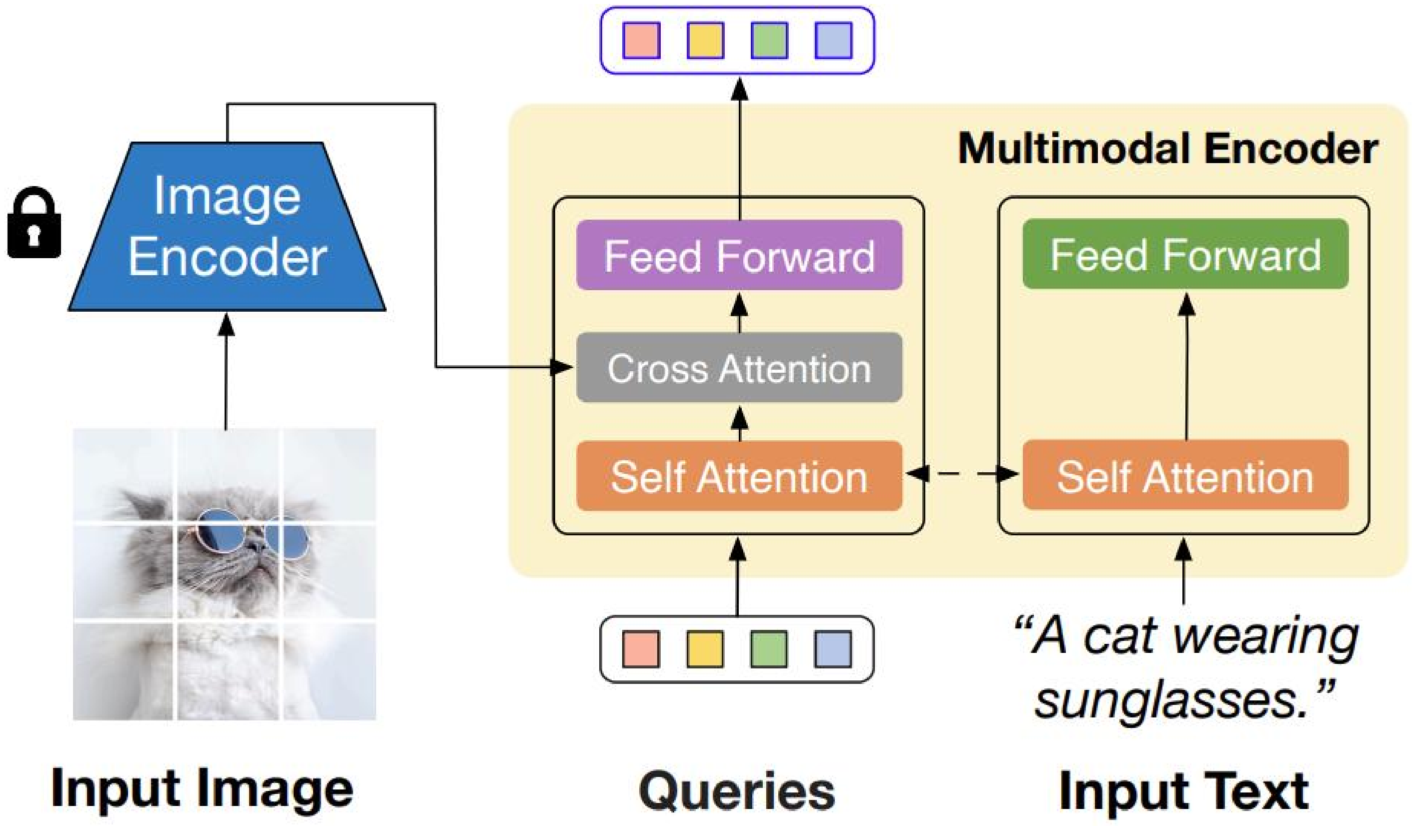
Background

Two stage pre-training



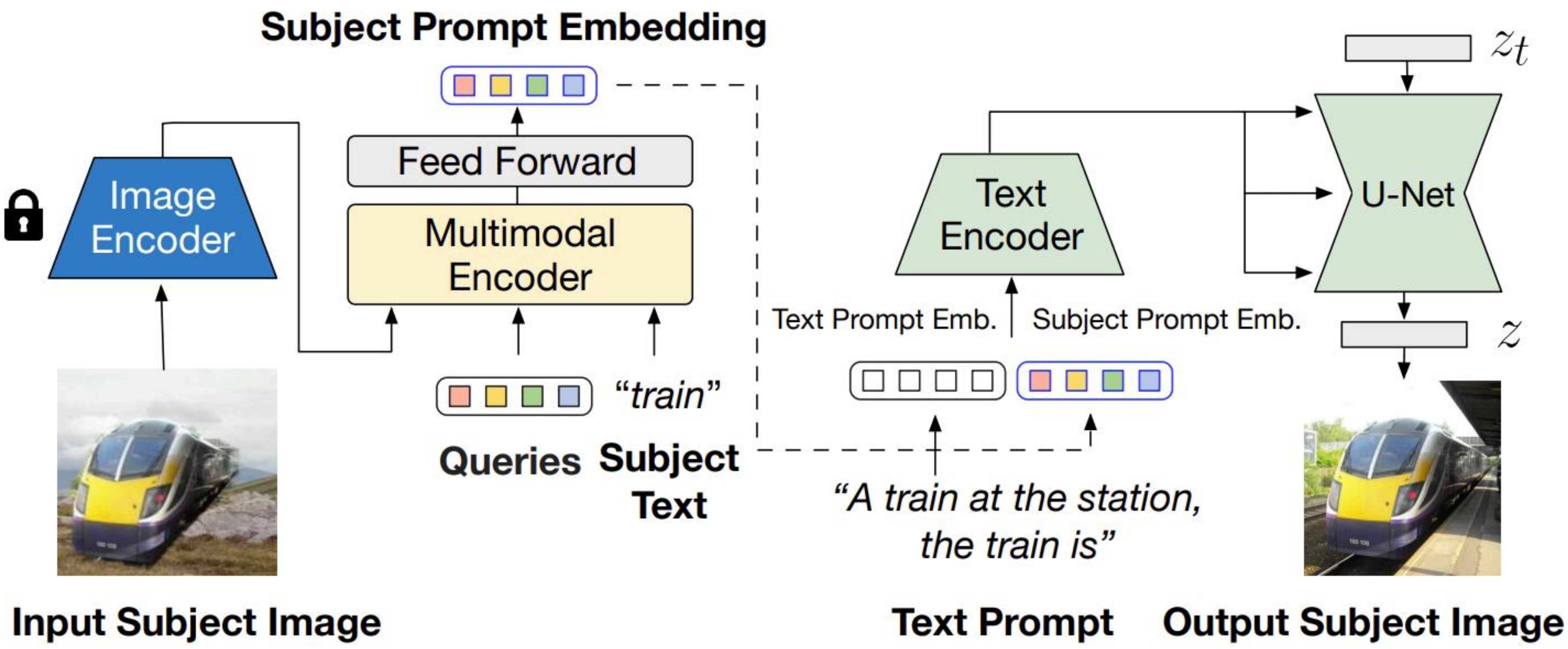
Background

BLIP2 Representation Learning Objectives



Background

BLIP2 Representation Learning Objectives



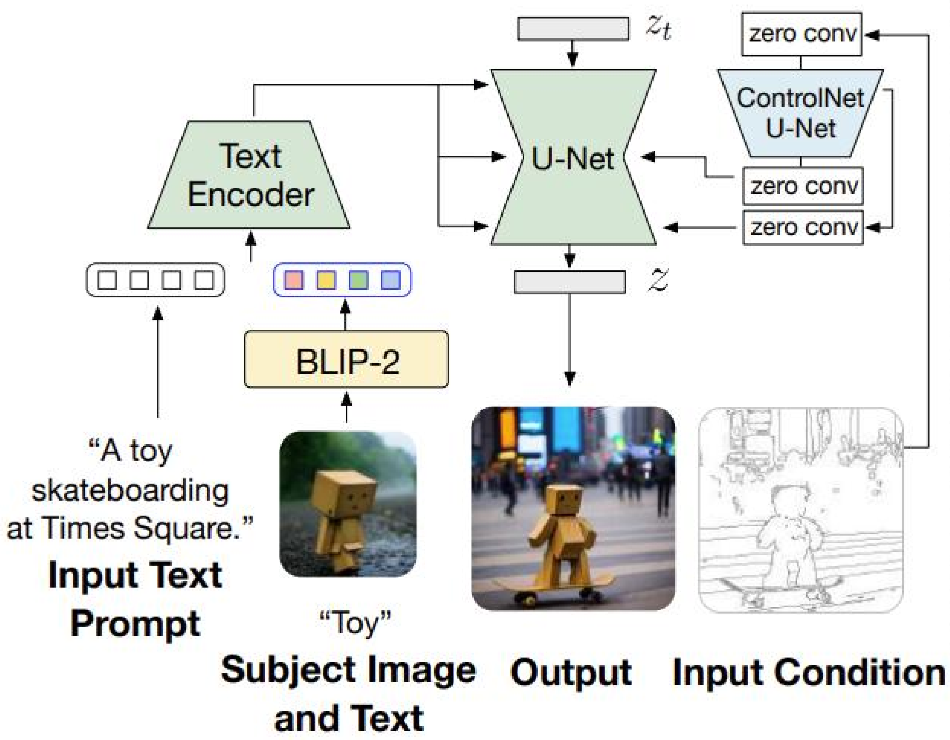
[text prompt], the [subject text] is [subject prompt]"

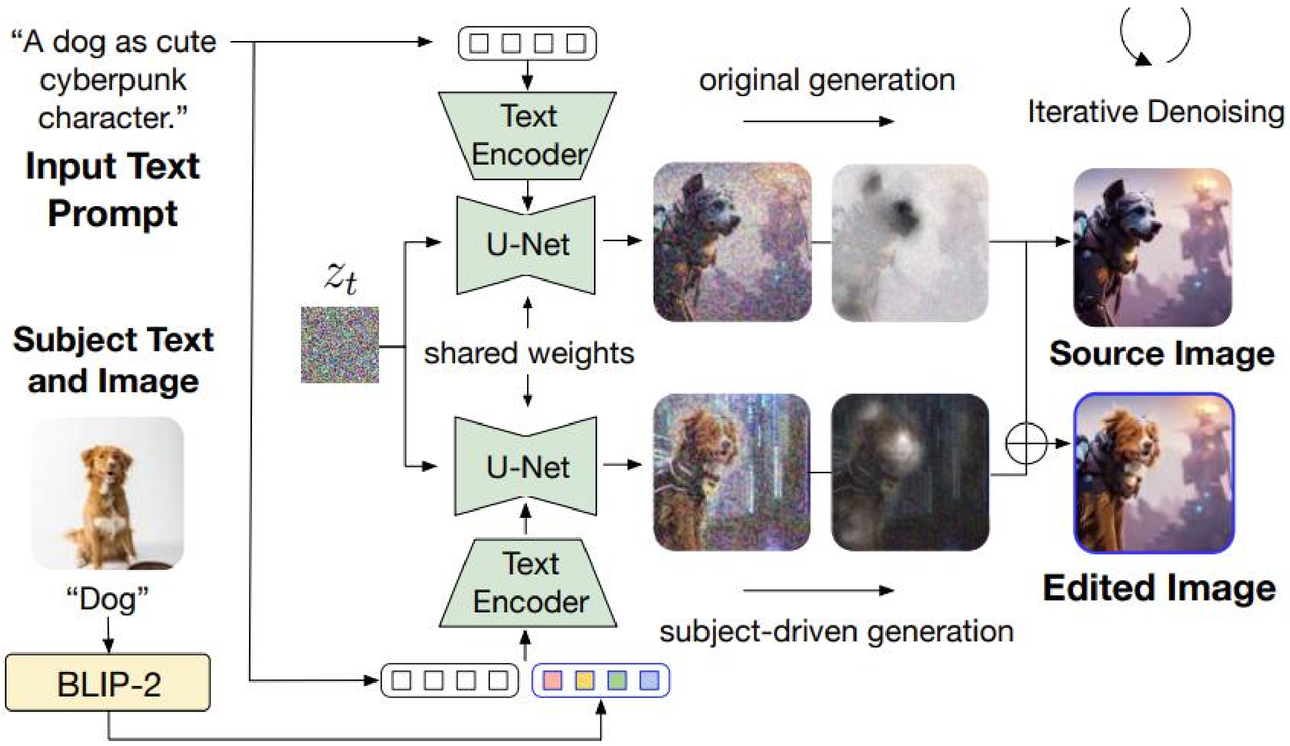
Background

Generating training image pair

Background

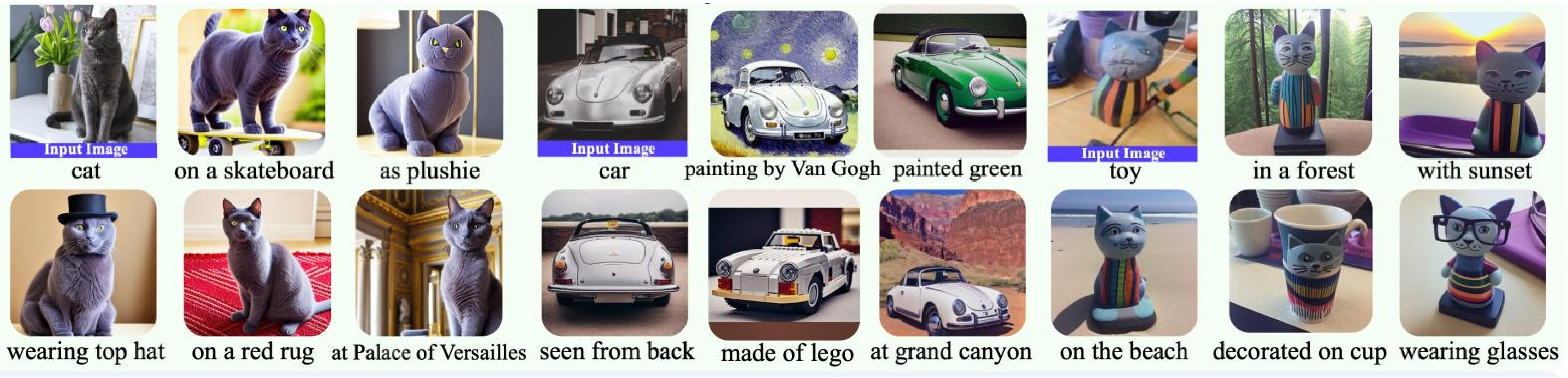
- Subject-specific Fine-tuning and Inference
- Structure-controlled Generation with ControlNet
- Subject-driven Editing with Attention Control





Background

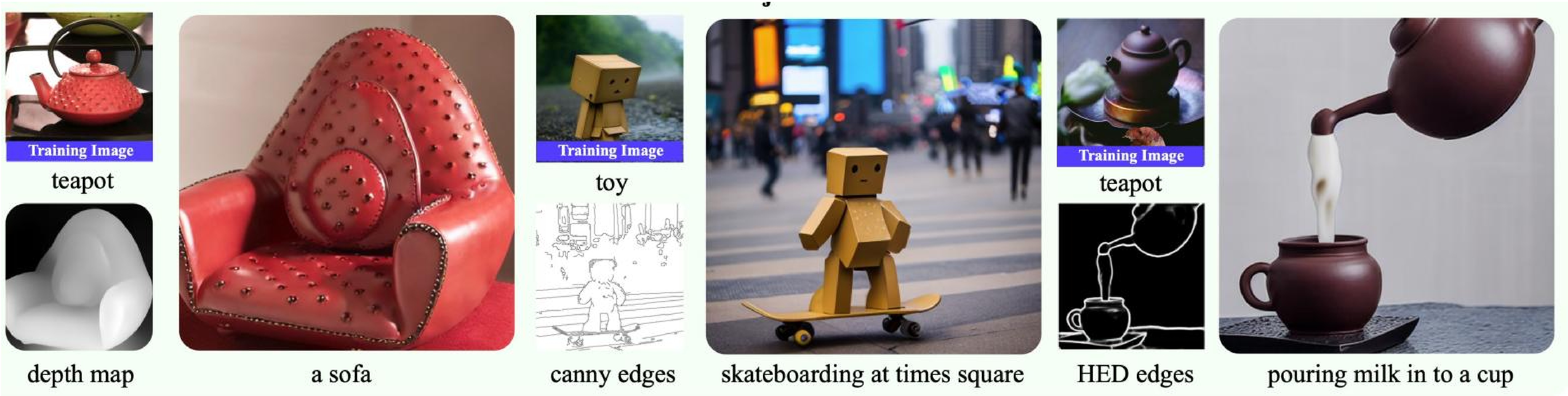
zero-sort subject-driven generation



Background few-step fine-tuned subject-driven generation

Background

structure controled subject-driven generation



Background

subject-driven image editing with attention control

Background

- context-appearence entanglement
- failing to address the text prompt
- wrong spatial composition

Subject images | A backpack on top of a **purple rug** in a forest. | Subject images | A **cube-shaped** bear plushie. | Subject images | A sneaker **on top of** a mirror.

Outline

Author

DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations

Tianhao Qi[1*], Shancheng Fang[1], Yanze Wu[2†], Hongtao Xie[1✉], Jiawei Liu[2],
Lang Chen[2], Qian He[2], Yongdong Zhang[1]
[1]University of Science and Technology of China, [2]ByteDance,
[*]Works done during the internship at ByteDance, [†]Project lead, [✉]Corresponding author,

Author

Tianhao Qi

PhD, University of Science and Technology of China
在 mail.ustc.edu.cn 的电子邮件经过验证
cross-modal generation object detection

First author: PhD student in University of Science and Technology of China, major in cross-modal generation, object detection

Author

Zhang Yongdong

University of Science and Technology of China
在 ustc.edu.cn 的电子邮件经过验证

1999-2002 Tianjin University, PhD, Signal and Information Processing
2002-2017 Researcher, Institute of Computing Technology, Chinese Academy of Sciences
Professor at the University of Science and Technology of China from 2017 to present

Research directions include multimedia content analysis, cybersecurity, and computational imaging.

Method

Results preview

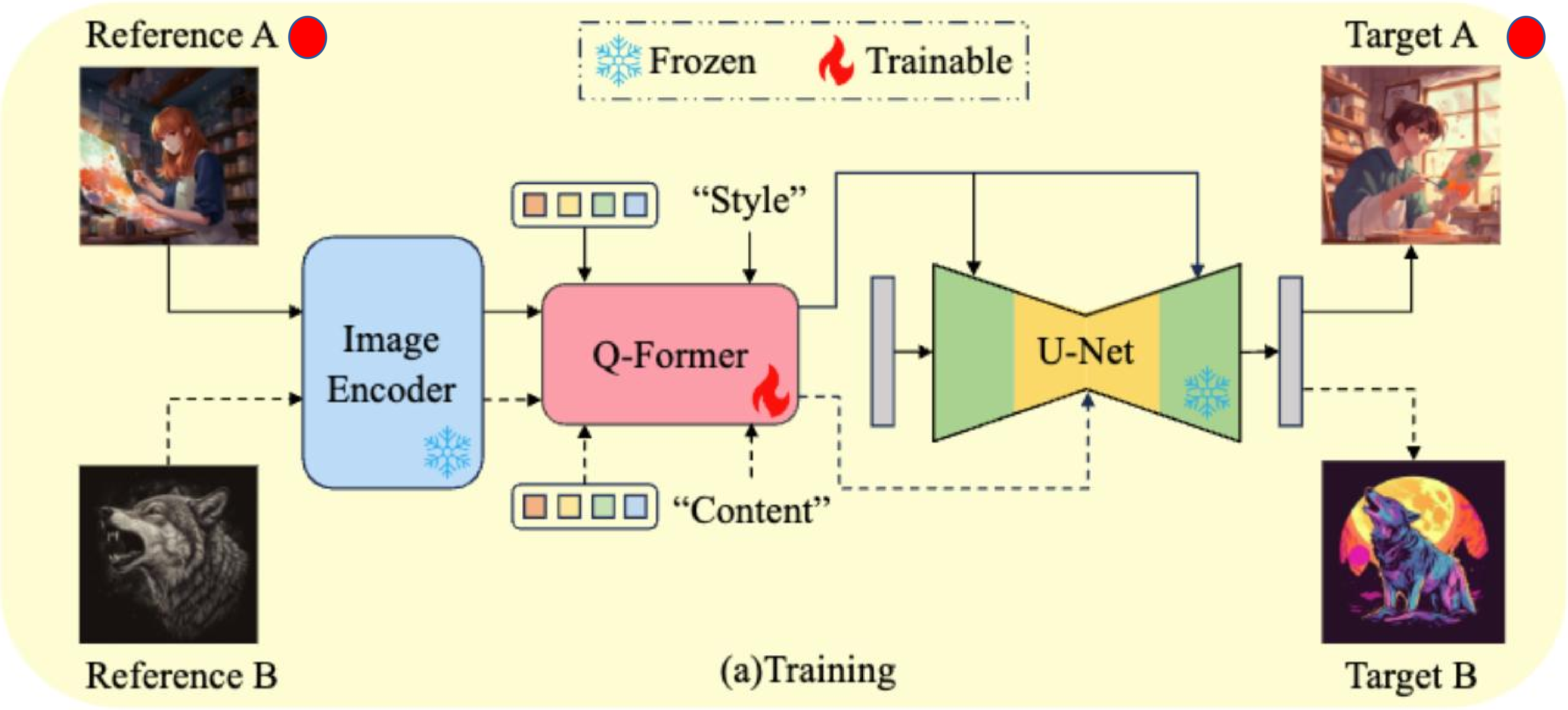
Method

Contributions

- Disentangle style and semantic representation of the reference image
- Injecting image style/semantic representation to different crossattention layers
- Established paired datasets

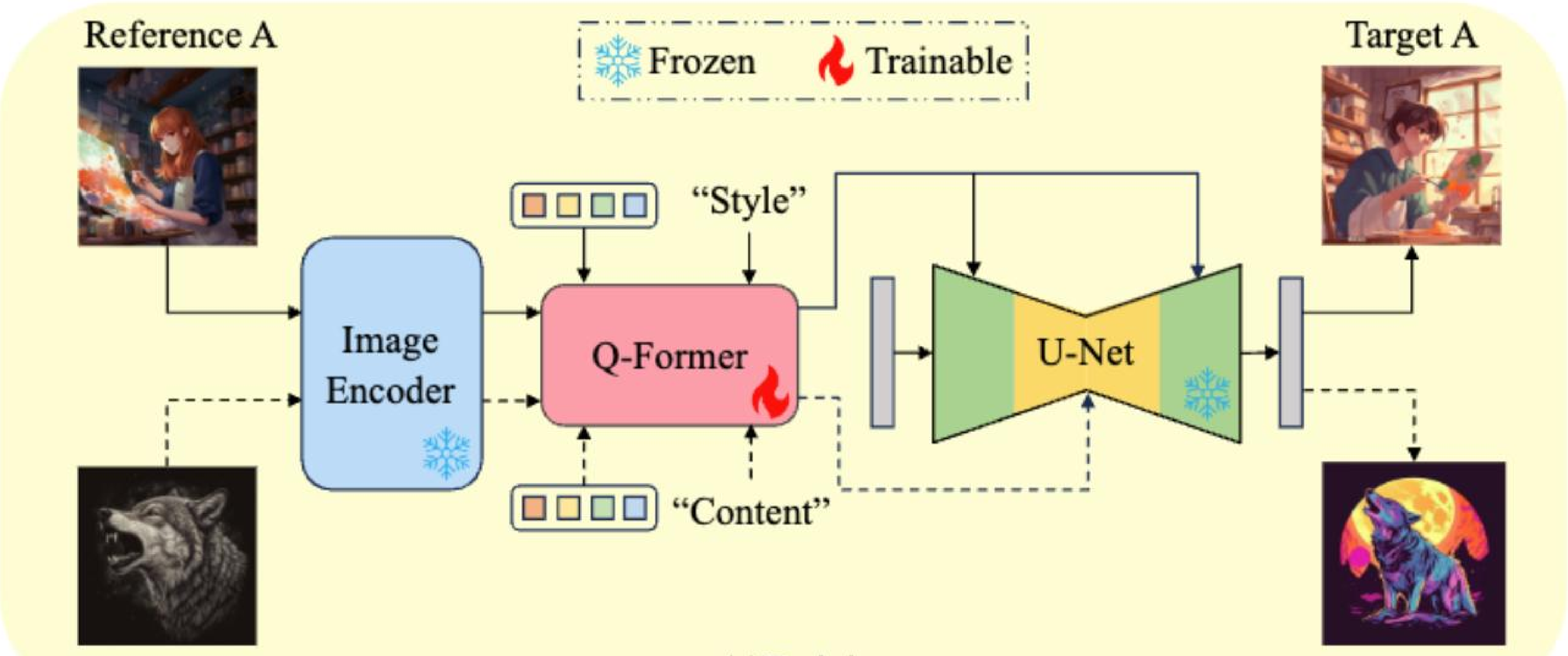
Method

- **Disentangle style and semantic representation of the reference image**



Method

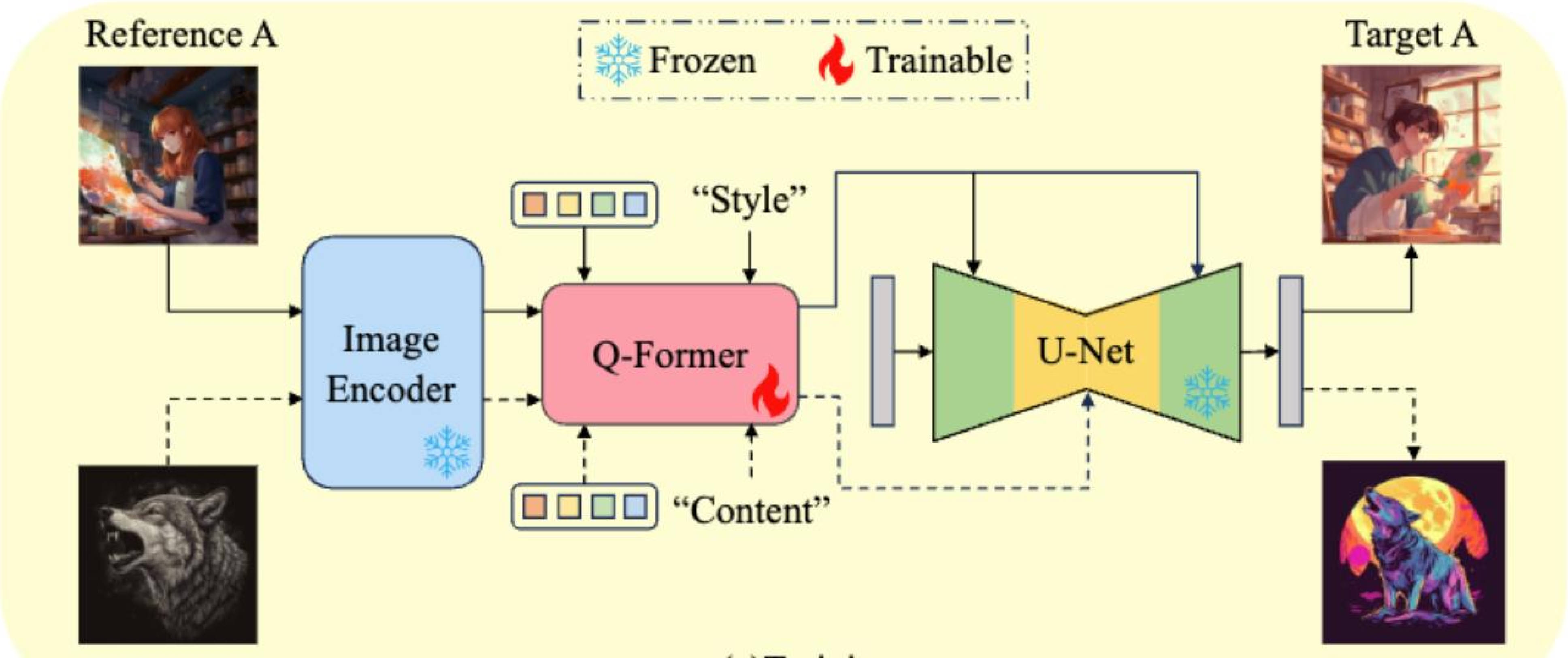
- Injecting image style/semantic representation to different



(a)Training

Method

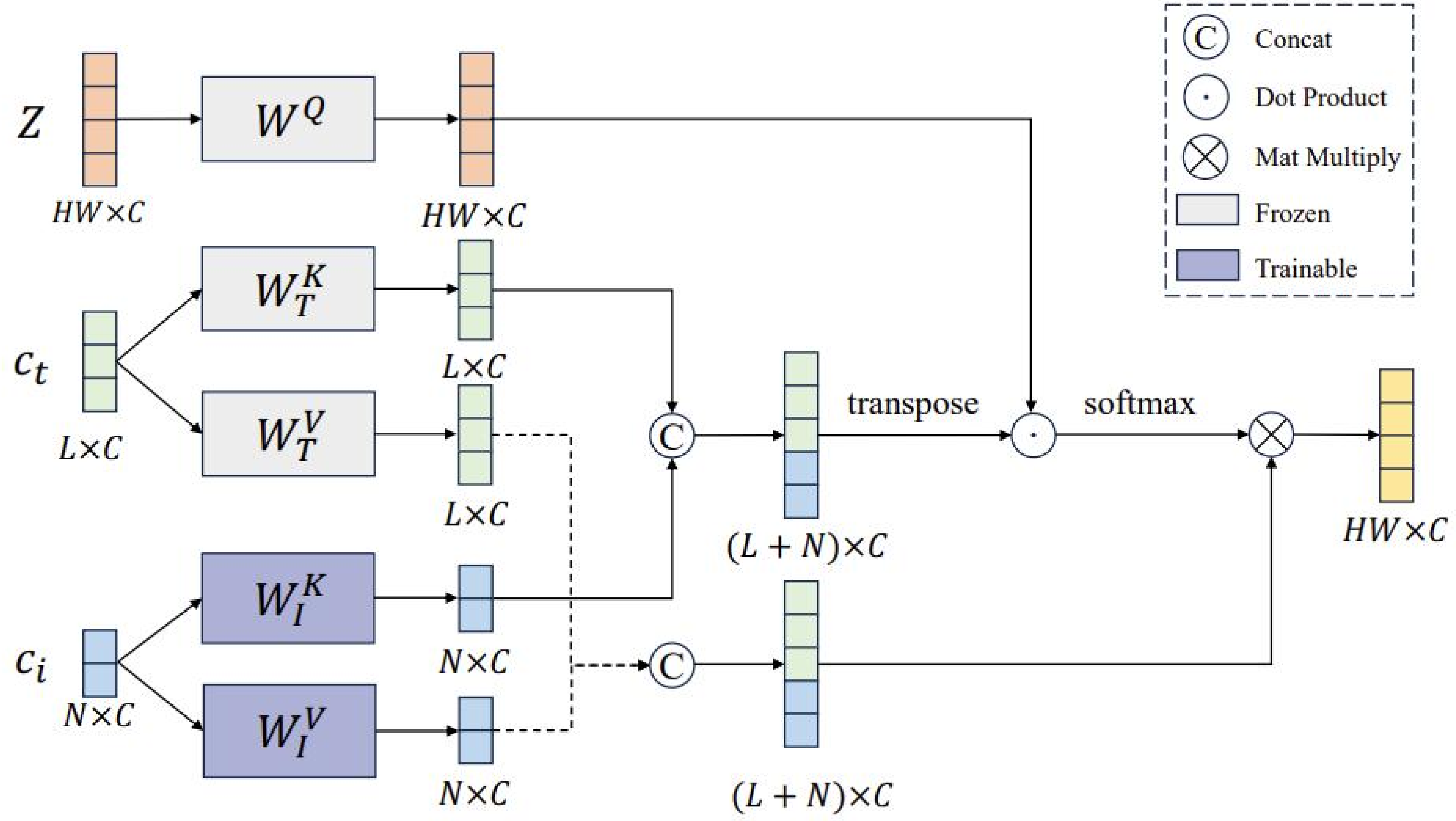
- **Injecting image style/semantic representation to different crossattention layers**



(a)Training

Method

- **Injecting image style/semantic representation to different crossattention layers**



$$Q = ZW^Q,$$

$$K = Concat(c_t W_T^K, c_i W_I^K),$$

$$V = Concat(c_t W_T^V, c_i W_I^V),$$

$$Z^{new} = Softmax(\frac{QK^T}{\sqrt{d}})V.$$

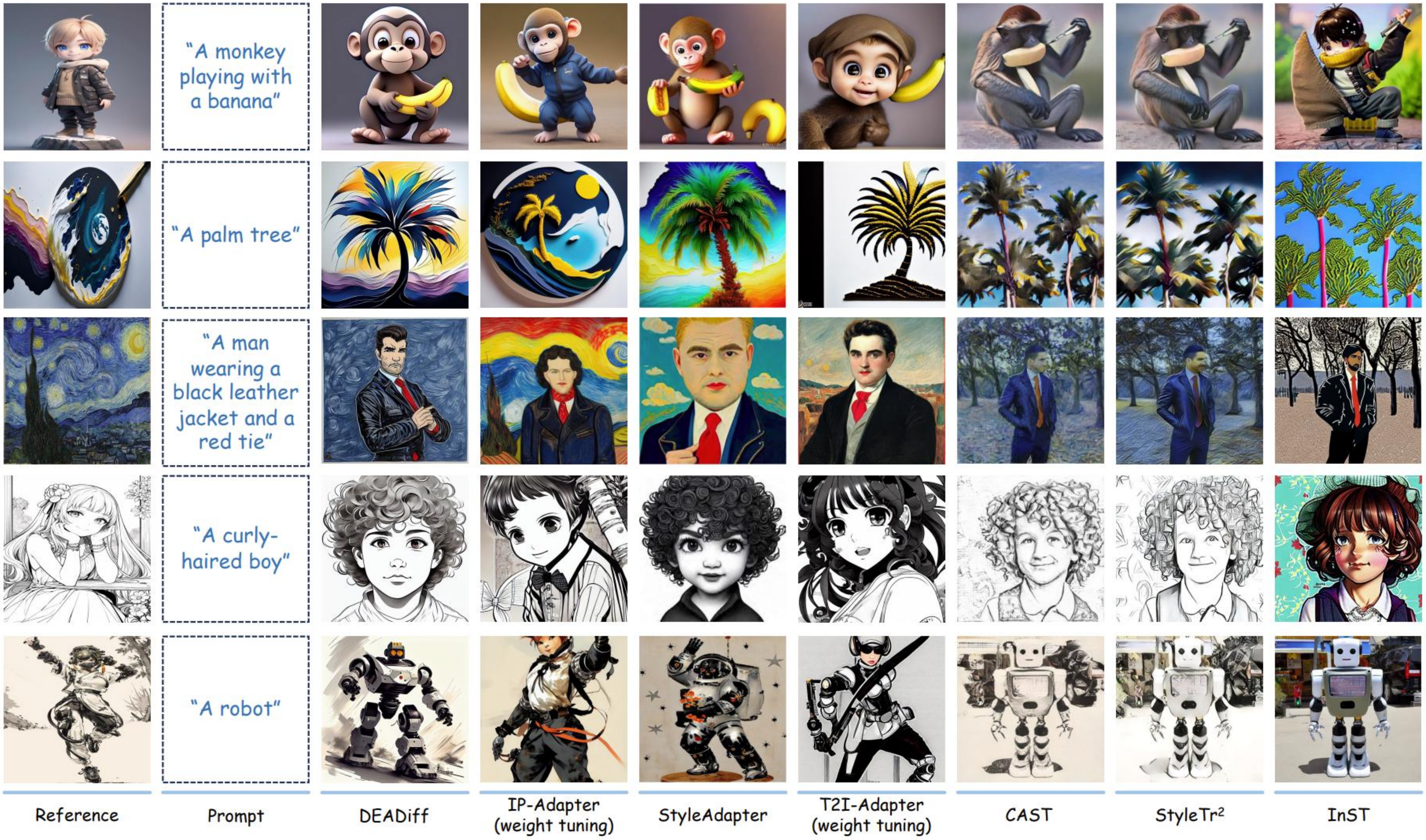
Method

- Establishing paired datasets
  - Text prompt combination
  - Image generation and collection
  - Paired images selection

Outline

Experiments

- Evaluation
  - Style Similarity (SS)
  - Text Alignment capability (TA)
  - Image Quality (IQ)
  - Subjective Preference (SP)


"A monkey playing with a banana"
"A palm tree"
"A man wearing a black leather jacket and a red tie"
"A curly-haired boy"
"A robot"
Reference
Prompt
DEADiff
IP-Adapter (weight tuning)
StyleAdapter
T2I-Adapter (weight tuning)
CAST
StyleTr²
InST

Experiments

Quantitative Comparisons

Method	SS↑	IQ↑	TA↑	SP↑
InST [37]	0.215	5.148	0.237	6.3
CAST [36]	0.224	4.922	<u>0.282</u>	8.7
StyTr^2 [3]	0.214	5.037	<u>0.282</u>	<u>13.1</u>
T2I-Adapter [17]	0.241	5.500	0.224	2.7
IP-Adapter [34]	**0.274**	<u>5.598</u>	0.155	-
DEADiff	<u>0.229</u>	**5.840**	**0.284**	**69.0**

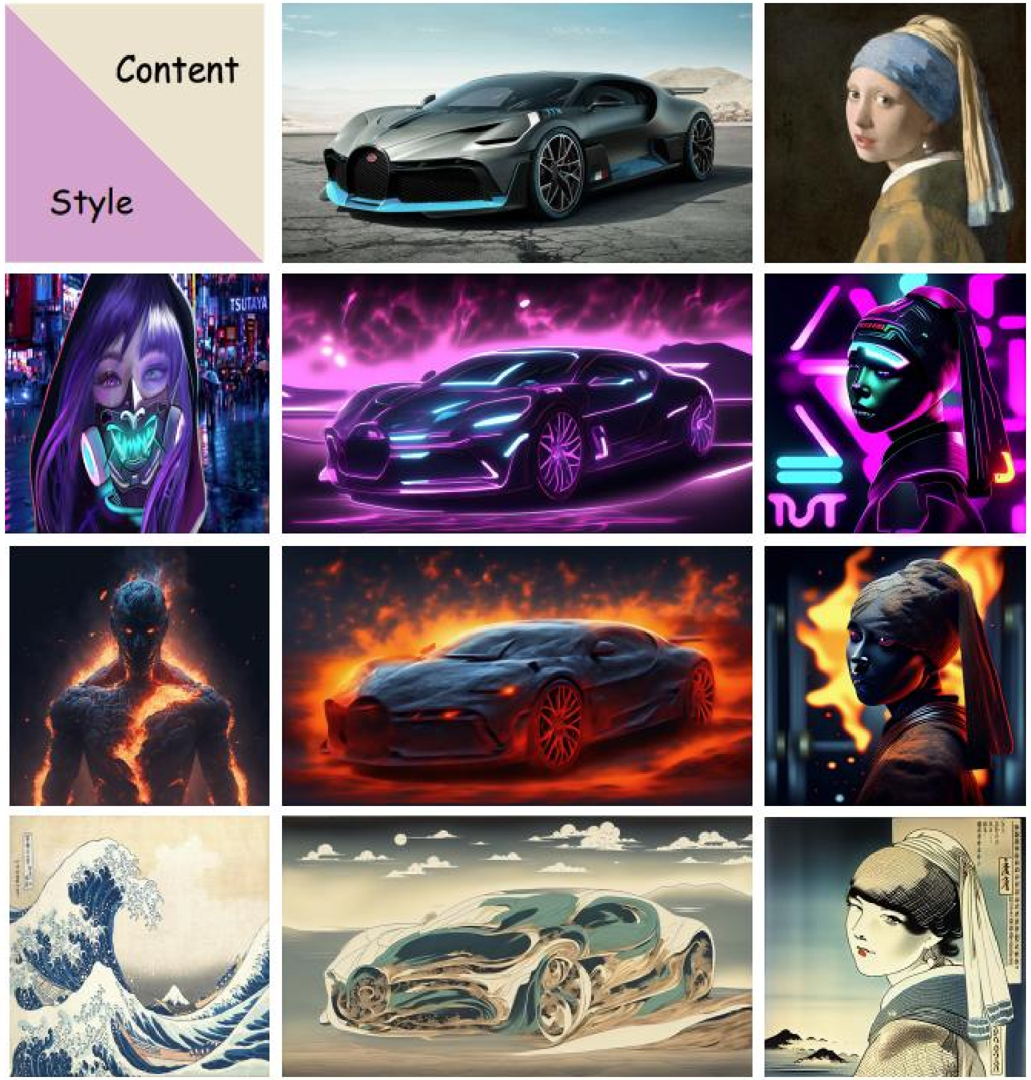
Experiments

Ablations

Method	Style Similarity↑	Text Alignment↑
Baseline	**0.274**	0.148
+ DCM	0.259	0.224
+ STRE	0.222	0.286
+ SERE	0.221	0.287
DEADiff	0.224	**0.289**

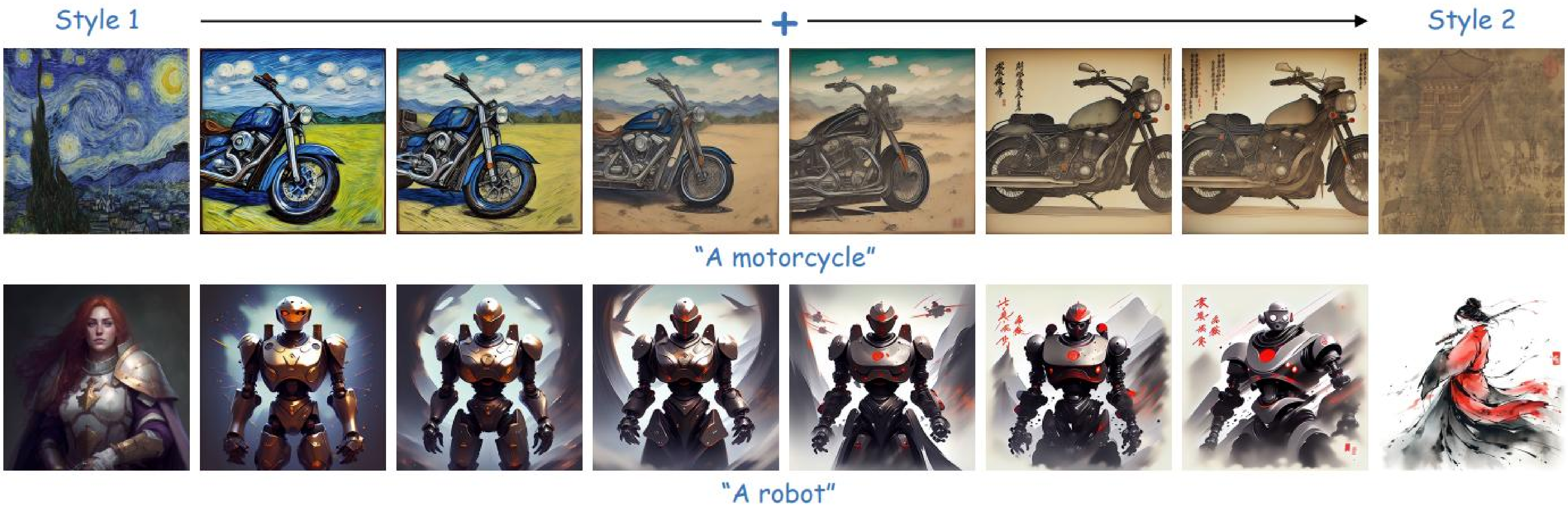
Experiments

Applications



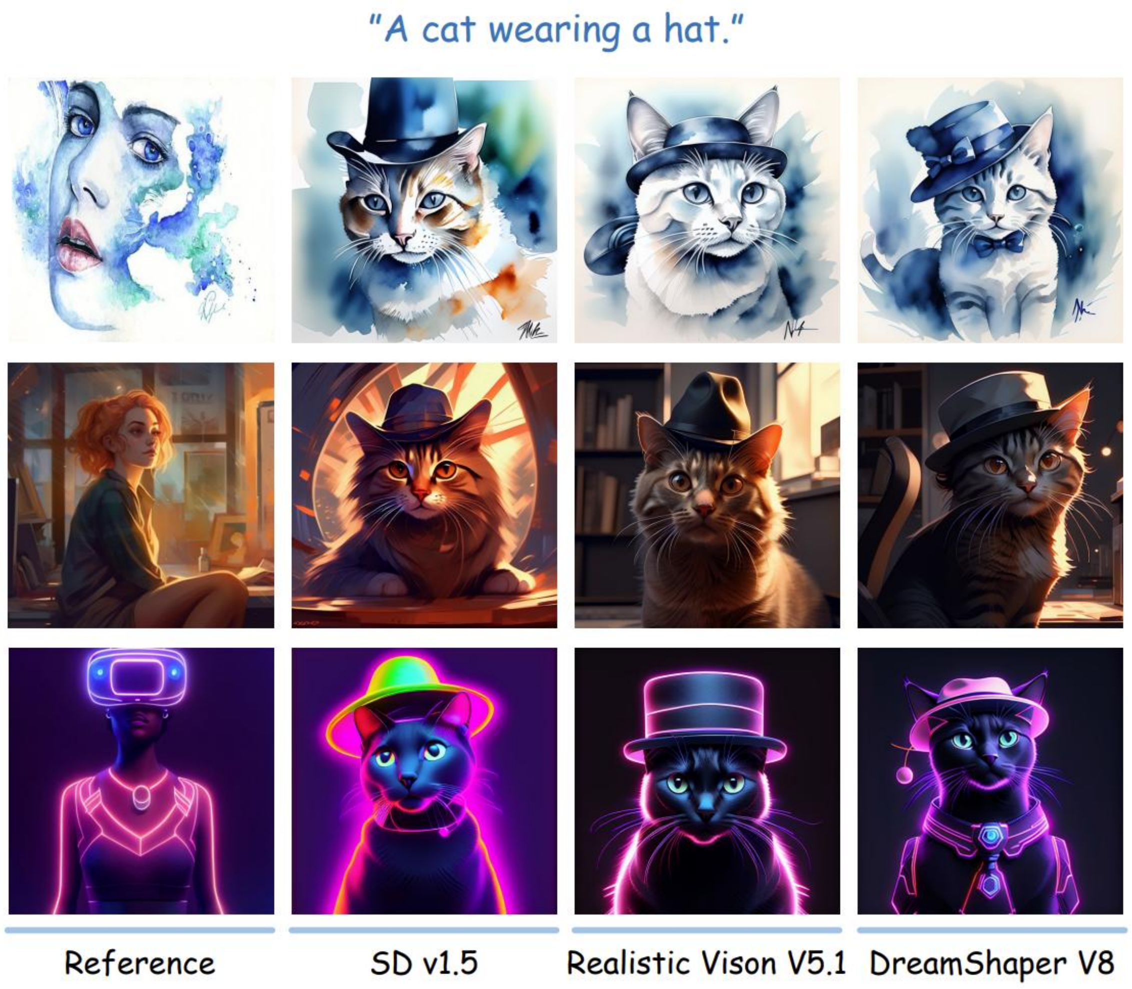
Experiments

Ablations



Experiments

Ablations



Thanks!

Background

BLIP2 pre-training