DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations

Tianhao Qi Shancheng Fang Yanze Wu Hongtao Xie Jiawei Liu Lang Chen Qian He Yongdong Zhang

University of Science and Technology of China, ByteDance

Introduction

Task Definition

Stylized Image Generation: Given a reference image, generate new images that imitate the style of the reference one and comply with additional text prompts simultaneously.

Research Issue

Previous encoder-based style transfer methods built upon diffusion models (*e.g.*, T2I-Adapter) introduce a particularly vexing issue: while they allow the model to follow the style of the reference image, they significantly diminish the model's performance in understanding the semantic context of text conditions.

Causing Factors of the Issue

- The extracted image features by encoders encompass both stylistic and semantic information, while the latter conflicts with the semantics in the text conditions.
- The reconstruction task adopted by previous methods encourages the diffusion model to excessively focus on the reference image, while neglecting the original text condition.

Contributions

- We propose a dual decoupling representation extraction (DDRE) mechanism to separately obtain style and semantic representations of the reference image, alleviating the problem of semantics conflict between text and reference images from the perspective of learning tasks.
- We introduce a disentangled conditioning mechanism that allows different parts of the cross-attention layers to be responsible for the injection of image style/semantic representation separately, reducing the semantics conflict further from the perspective of model structure.
- We build two paired datasets to aid the DDRE mechanism using the non-reconstruction training paradigm.

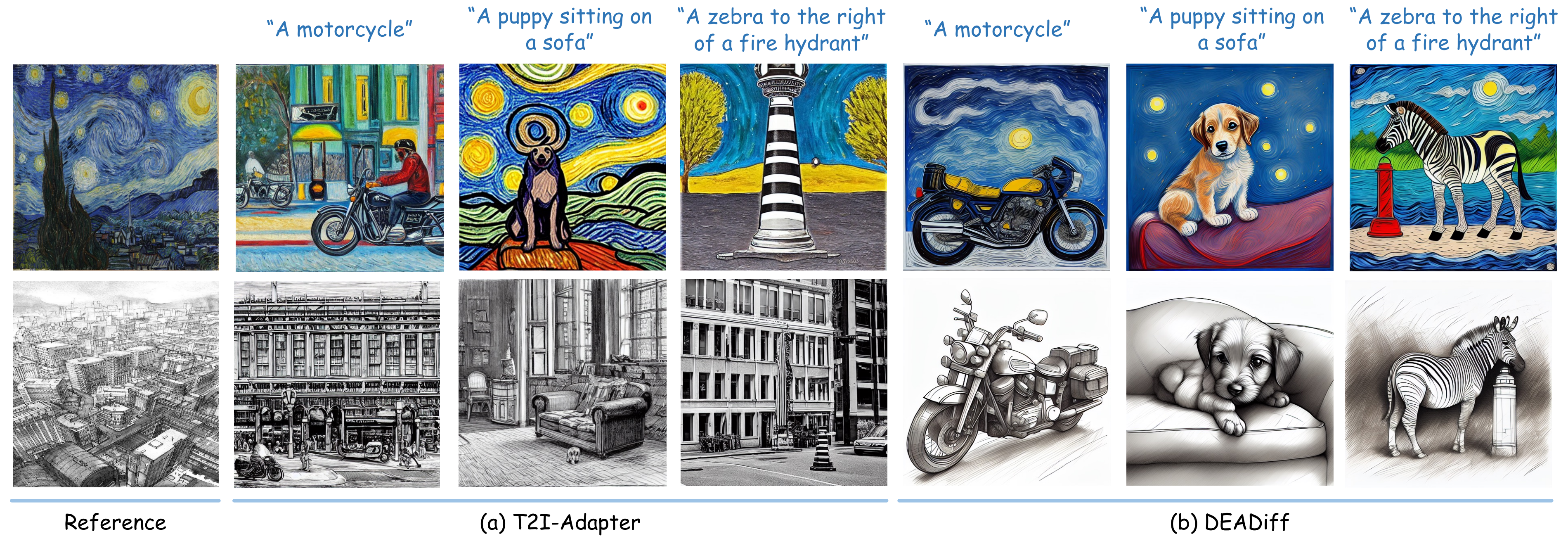

"A motorcycle"
"A puppy sitting on a sofa"
"A zebra to the right of a fire hydrant"
"A motorcycle"
"A puppy sitting on a sofa"
"A zebra to the right of a fire hydrant"
Reference
(a) T2I-Adapter
(b) DEADiff

Figure 1. Given a style reference image, **DEADiff** is capable of synthesizing new images that resemble the style and are faithful to text prompts simultaneously. However, previous encoder-based methods (*i.e.*, T2I-Adapter) significantly impair the text controllability of the diffusion-based text-to-image models.

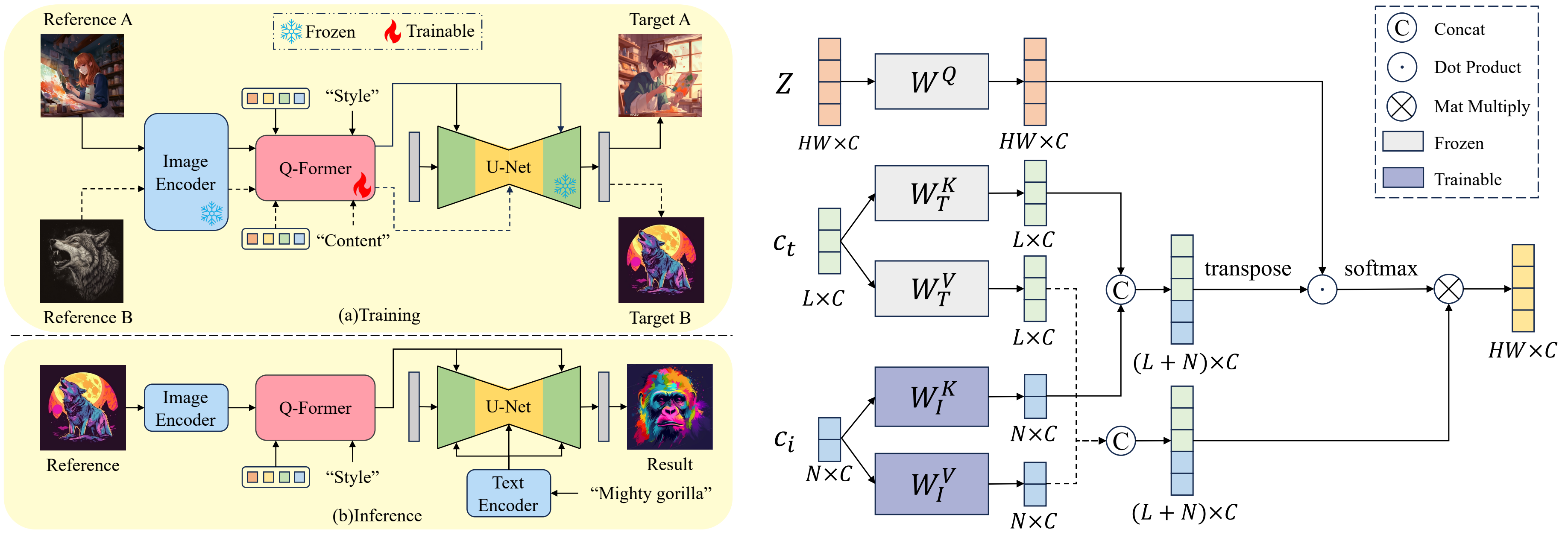
Method

Core Idea

To remove the interference of semantics from the reference image, only extract its style-relevant features and inject them into the layers responsible for capturing style features in the diffusion U-Net, as illustrated by Fig. 2a.

Dual Decoupling Representation Extraction

- **STyle Representation Extraction (STRE)**: Sample a pair of distinct images, both maintaining the same style but serving as the reference and target respectively. Instruct the Q-former with the "style" condition to extract style-relevant features of the reference image by the learnable query tokens. Then the output features are provided for conditioning to the denoising U-Net coupled with the caption detailing the content of the target image.
- **SEmantic Representation Extraction (SERE)**: Select two images that share the same subject matter but exhibit distinct styles, which are assigned as the reference and target images. Instruct the Q-former with the "content" condition to extract content-specific representations of the reference image. Then supply the output and the text style words of the target image, concurrently, as the conditioning for the denoising U-Net.
- **Reconstruction Task**: To prevent information loss, combine the output of the query tokens of the Q-former under "style" and "content" conditions as the conditioning for the U-Net to reconstruct the same image.



(a) The training and inference paradigm of **DEADiff**. We use proprietary paired datasets for training Q-Former to extract disentangled representations under conditions "style" and "content", which are injected into mutually exclusive cross-attention layers. **Only the parameters of the Q-former and extra linear projection layers for processing image features are updated during training.**

(b) The illustration of our proposed joint text-image cross-attention layer. We add two trainable linear projection layers W_I^K, W_I^V to process image features c_i, in conjunction with frozen ones W_T^K, W_T^V for text features c_t. We concatenate the key and value matrices from text and image features respectively, subsequently initiating a single cross-attention operation with U-Net query features Z.

Disentangled Conditioning Mechanism

- Conditions the coarse cross-attention layers with lower spatial resolution on semantics, while the fine cross-attention layers with higher spatial resolution are conditioned on the style.
- Devise a joint text-image cross-attention layer (Fig. 2b) to make the U-Net support image features as conditions.

$$Q = ZW^Q, \tag{1}$$
$$K = Concat(c_tW_T^K, c_iW_I^K), \tag{2}$$
$$V = Concat(c_tW_T^V, c_iW_I^V), \tag{3}$$
$$Z^{new} = Softmax(\frac{QK^T}{\sqrt{d}})V. \tag{4}$$

Paired Datasets Construction

- **Step 1**: Combine the subject words and the style words to form text prompts.
- **Step 2**: Generate images by Midjourney under the guidance of text prompts and collect the results.
- **Step 3**: Construct two paired datasets by selecting images with the same style words, and the same subject words, respectively.

Training and Inference

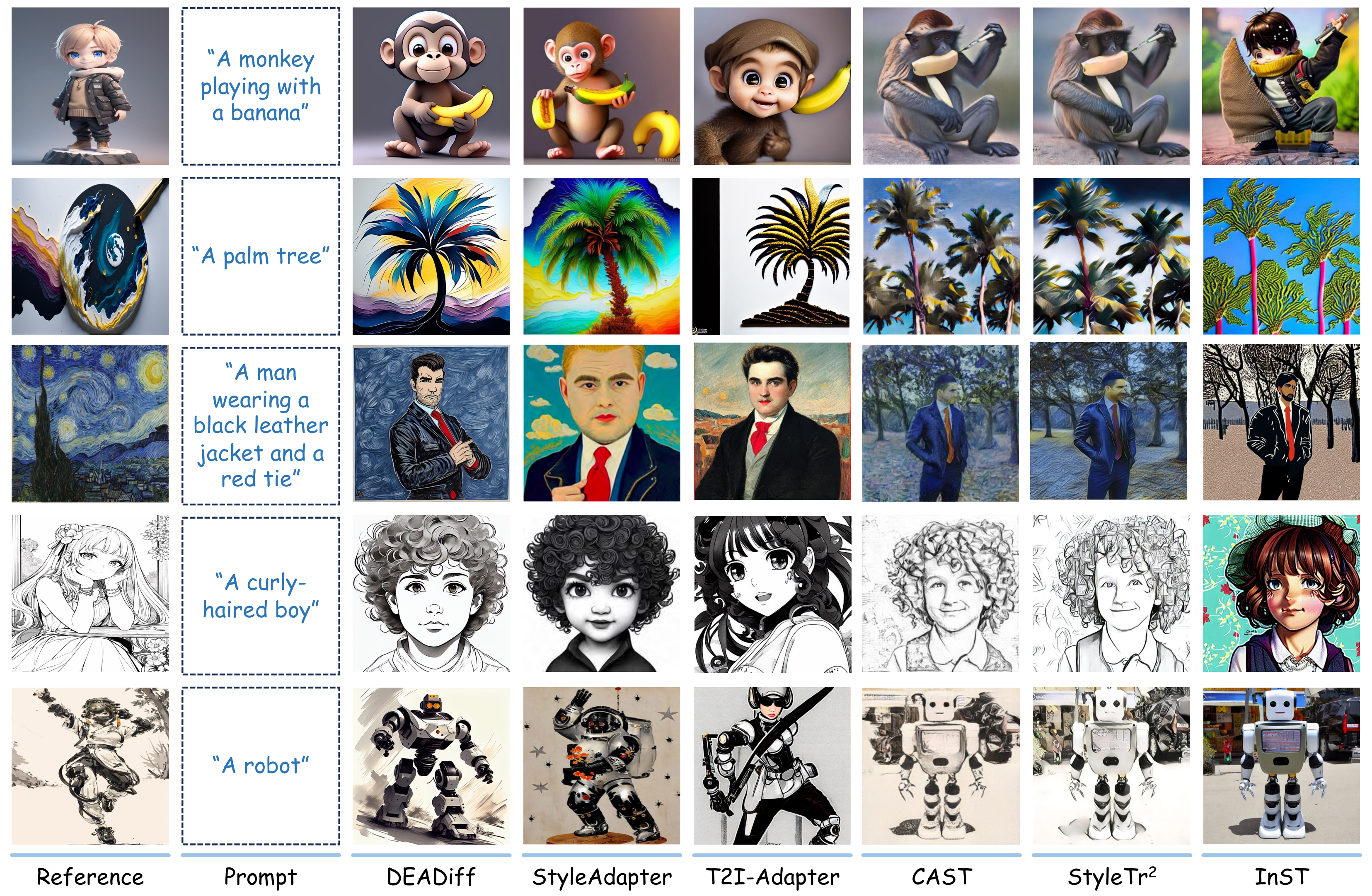
- Training loss:

$$L = \mathbb{E}_{z,c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta\left(z_t,t,c\right)\|_2^2\right], \tag{5}$$

- **Inference paradigm**: Inject the output queries of the Q-Former with "style" conditions to fine layers of the diffusion U-Net, which respond to style information rather than global semantics.

Results

Comparisons with State-of-the-Arts

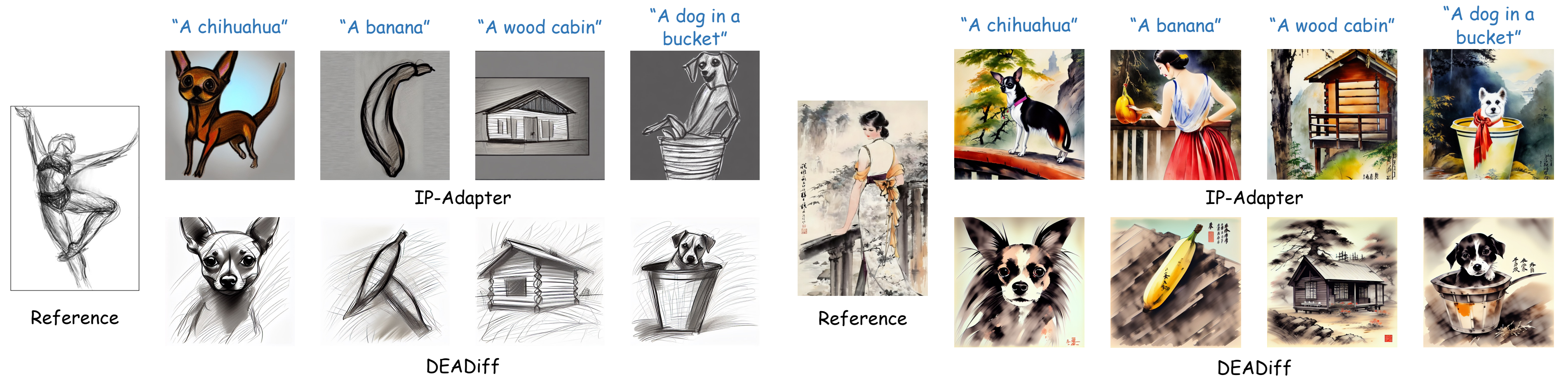


Qualitatively: Given a reference image and a text prompt, **DEADiff** not only better adheres to the textual prompts but also significantly preserves the style and detailed textures of the reference image, with very minor differences in the color tones.

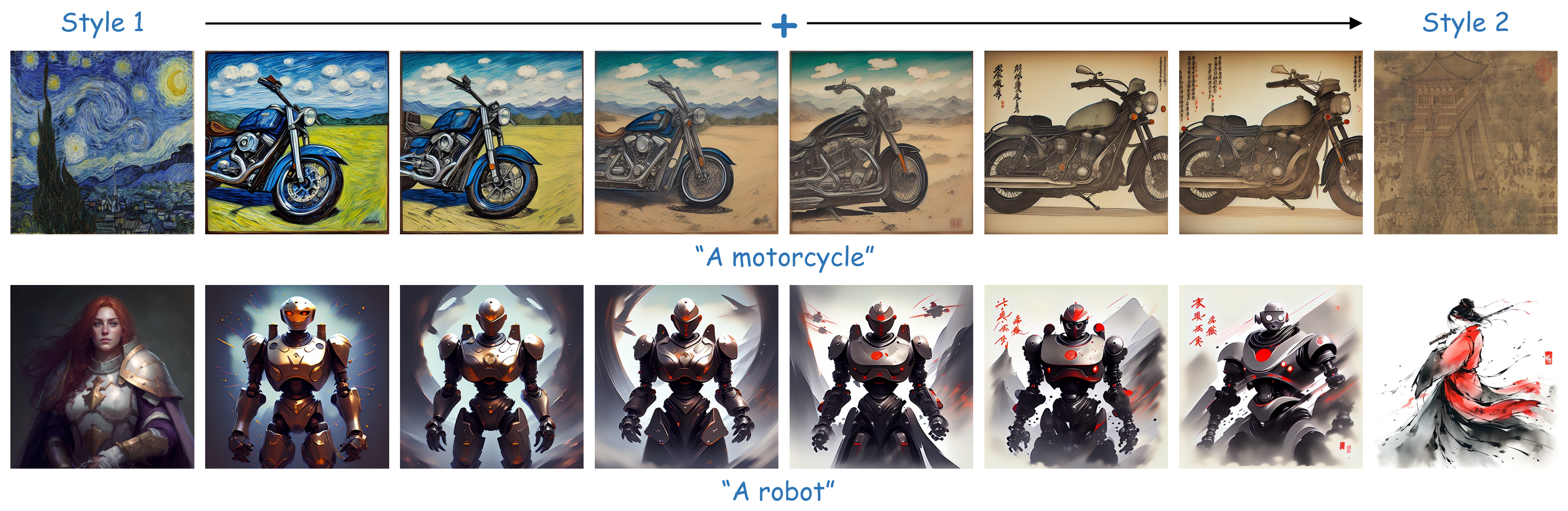
Method	Style Similarity↑	Image Quality↑	Text Alignment↑	User Preference↑
InST	0.215	5.148	0.237	6.3
CAST	0.224	4.922	0.282	8.7
StyTr^2	0.214	5.037	0.282	13.1
T2I-Adapter	**0.241**	5.500	0.224	2.7
DEADiff	0.229	**5.840**	**0.284**	**69.0**

Quantitatively: **DEADiff** achieves an optimal balance between text fidelity and style similarity with the most pleasing image quality. Meanwhile, the big margin in user preference achieved by **DEADiff** further demonstrates its broad application prospects.

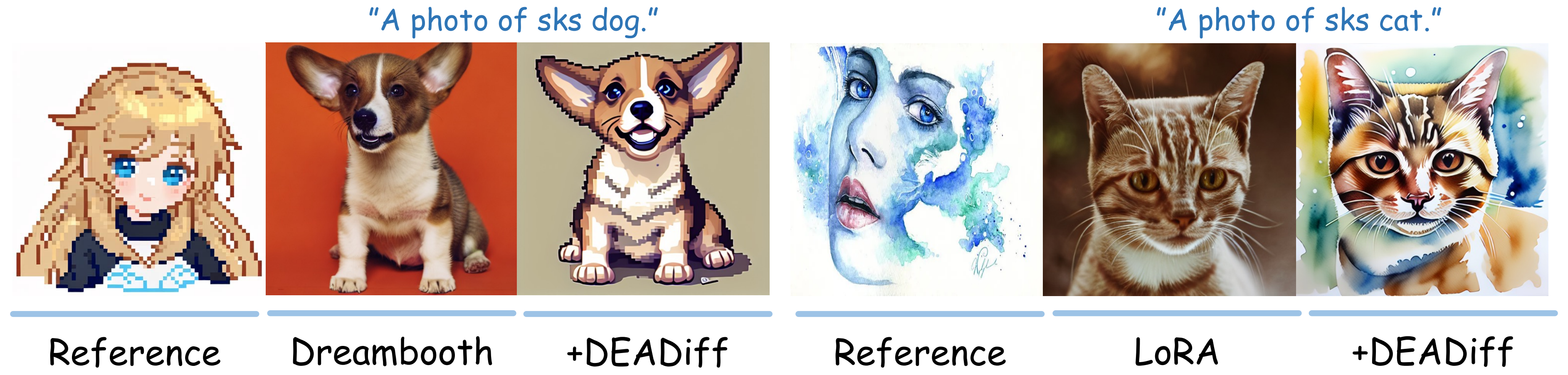
Comparisons with IP-Adapter



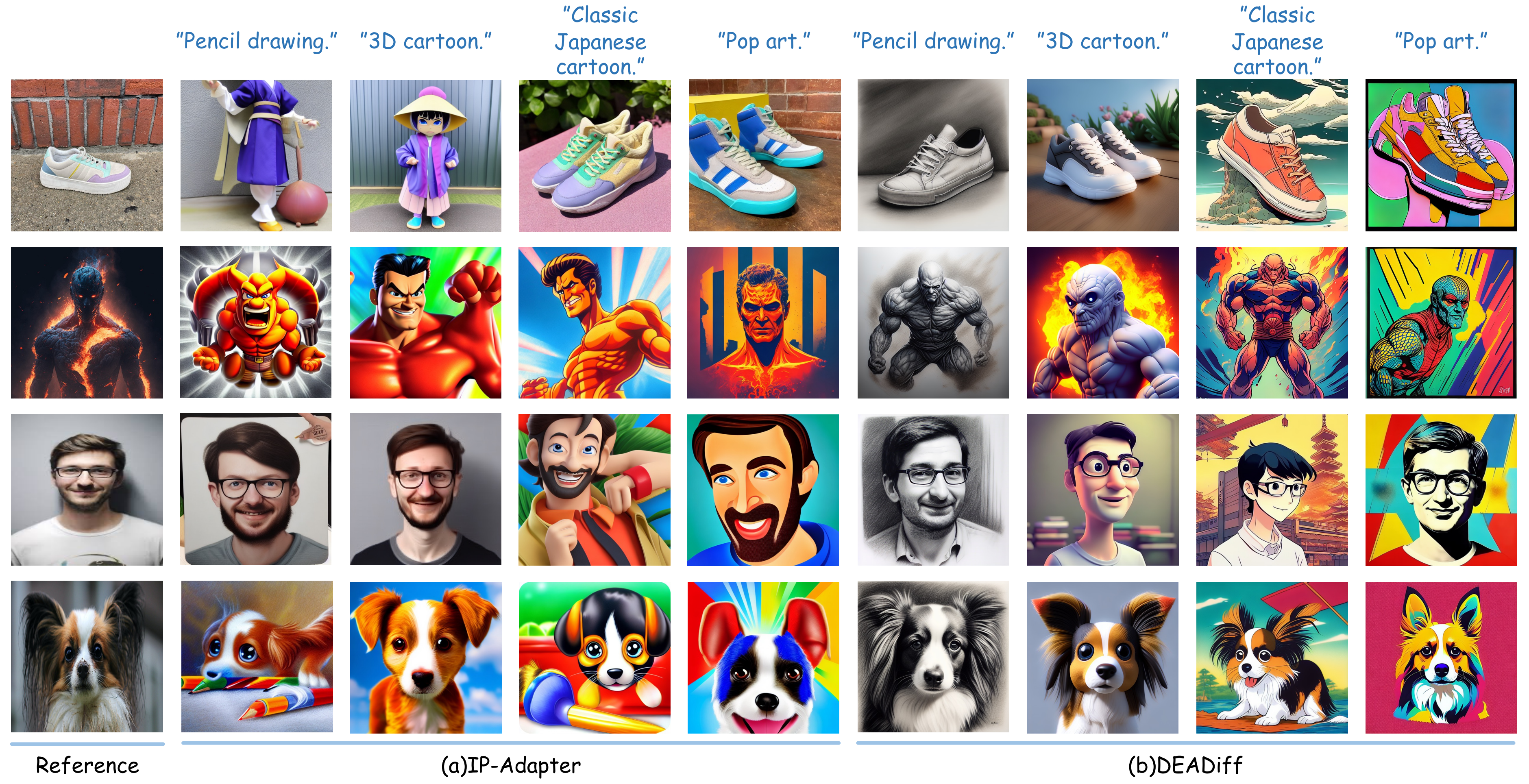
Style Mixing



DreamBooth/LoRA + DEADiff



Stylized Reference Object Generation



This is a novel application of **DEADiff**. Instructed by the "content" condition, **DEADiff** can extract the semantics of a reference image and achieve its stylization under the guidance of an extra text prompt.

Conclusion & Contact

We delve into the reasons for the decline in text control capabilities of existing encoder-based stylized diffusion models and subsequently propose the targeted design of **DEADiff**. It includes a dual decoupling representation extraction mechanism and a disentangled conditioning mechanism. Empirical evidence demonstrates that **DEADiff** is capable of attaining an optimal equilibrium between stylization capabilities and text control. Future work could aim to further enhance style similarity and decouple instance-level semantic information.

(a) Project Page (b) Arxiv (c) Code (d) Laboratory (e) Personal Page