DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations
Supplementary Material

7. Supplementary

7.1. Quantitative Comparisons

Given that *DEADiff* is proposed specifically to address the issue of text controllability loss inherent in encoder-based methods, we primarily emphasize the quantitative metric of text alignment in the main paper. Below, we additionally provide a quantitative comparison of the style similarity and image quality between *DEADiff* and the state-of-the-art methods, as illustrated by Tab. 3.

Evaluation Metrics.

Style Similarity: We propose a more reasonable approach to measure style similarity. Specifically, the procedure begins with using the CLIP Interrogator [2] to generate the optimal text prompts that align with the reference image. Subsequently, we filter out the prompts related to the content of the reference image and compute the cosine similarity between the remaining prompts and the generated image within the CLIP text-image embedding space. The computational result denotes the style similarity, effectively mitigating interference from the content of the reference image.

Image Quality: We adopt a prediction model named LAION-Aesthetics Predictor V2 [3] to assess the quality of images generated by each method.

Text Alignment: We determine the cosine similarity within the CLIP text-image embedding space between the textual prompts and their corresponding synthesized images, indicative of the text alignment capability.

Differing from Tab. 1, we not only list the quantitative results of T2I-Adapter [17] at the default image condition weight of 1.0, but also provide the results when the image condition weight is set to 0.9 and 0.8 in Tab. 3. Evidently, T2I-Adapter, under different image condition weights, exhibits a clear trade-off between style similarity and text alignment. When the image condition weight is overly large, *e.g.*, 1.0, the generated image essentially becomes a reorganization of the reference image. This leads to a high style similarity (0.241) but significantly weakens text controllability (0.224), as introduced in Sec. 1. However, if we reduce this weight, the style similarity will drop rapidly to 0.184. Fig. 11 provides an intuitive illustration that *DEADiff* is situated outside T2I-Adapter's trade-off curve, thereby demonstrating its enhanced ability to strike a balance between style similarity and text control capability. Moreover, *DEADiff* outperforms other top-performing methods in both style similarity and text alignment, including CAST, StyTr^2 and InST, further confirming the effectiveness of our approach. Meanwhile, the substantial advantage reflected in the image quality metric compared to all other methods corroborates the practicality of our approach.

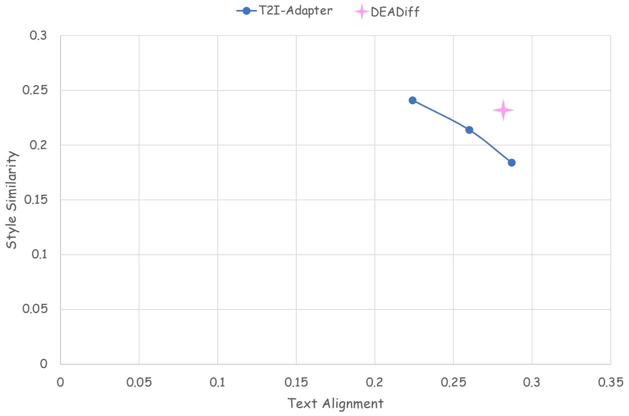

Figure 11. Quantitative comparison between *DEADiff* and the trade-off curve of T2I-Adapter.

7.2. User Study

In addition to objective evaluations, we have also designed a user study to subjectively assess the practical performance of various methods. Given 18 style reference images from Civitai [4], we employed CAST [36], InST [37], StyTr^2 [3], T2I-Adapter [17], and *DEADiff* to separately generate corresponding stylized results. Specifically, we utilized a total of 21 distinct text prompts. Thus, apart from three reference images corresponding to two prompts each, the remaining 15 reference images and 15 prompts are directly matched one-to-one. We asked 24 users from diverse backgrounds to evaluate the generated results in terms of text-image alignment, image quality, and style similarity, and to provide their overall preference considering these three aspects. Consequently, we have obtained a total of 2016 voting results. The final results are displayed in Tab. 4. *DEADiff* outperforms all state-of-the-art methods on three evaluation aspects and the overall preference with a big margin, which demonstrates the broad application prospects of our method.

7.3. Inference Efficiency

Despite *DEADiff* adding 1900 MB to the memory occupation, the increase in average inference time on one A100-80G GPU is only marginal, as shown in Tab. 5.

[2] https://github.com/pharmapsychotic/clip-interrogator

[3] https://github.com/christophschuhmann/improved-aesthetic-predictor

[4] https://civitai.com

Method	Style Similarity↑	Image Quality↑	Text Alignment↑
InST [37]	0.215	5.148	0.237
CAST [36]	0.224	4.922	0.282
StyTr^2 [3]	0.214	5.037	0.282
T2I-Adapter 1.0 [17]	**0.241**	5.500	0.224
T2I-Adapter 0.9 [17]	0.214	5.534	0.260
T2I-Adapter 0.8 [17]	0.184	5.580	**0.287**
DEADiff	0.229	**5.840**	0.284

Table 3. Quantitative comparison of style similarity, image quality and text alignment with the state-of-the-art methods. **Bold** numbers denote the best results, while the underlined numbers denote the second best results. We show different results for T2I-Adapter with three varying condition weights: 1.0, 0.9, and 0.8, which presents an obvious trade-off between style similarity and text alignment.

Aspect	Style Similarity↑	Image Quality↑	Text Alignment↑	Overall↑
InST [37]	7.8	8.5	11.9	6.3
CAST [36]	8.7	9.3	10.5	8.7
StyTr^2 [3]	16.1	11.5	13.9	13.1
T2I-Adapter [17]	1.9	8.1	7.5	2.7
DEADiff	**65.4**	**62.5**	**56.2**	**69.0**

Table 4. Results for the user study in percentages.

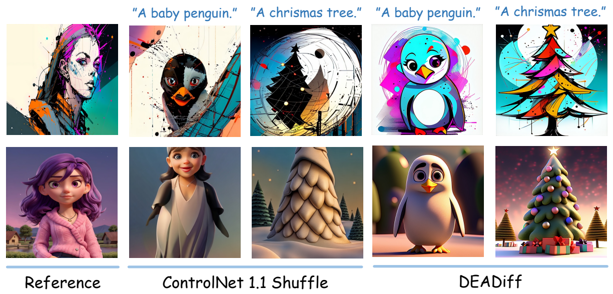

Figure 12. Visual comparison between ControlNet 1.1 Shuffle and *DEADiff*.

Model	SD	ControlNet 1.1 Shuffle	*DEADiff*
Memory (MB)	7774	10986	9674
50-Step DDIM Time on A100 (s)	2.28	3.00	2.43

Table 5. Memory usage and sampling time on 1 A100-80G GPU.

7.4. Comparison with ControlNet 1.1 Shuffle

We compare our method with ControlNet 1.1 Shuffle [5] and present the results in Fig. 12. It is clear that our method outperforms ControlNet 1.1 Shuffle in carving the style of the reference image, fidelity to the text, and generated image quality.

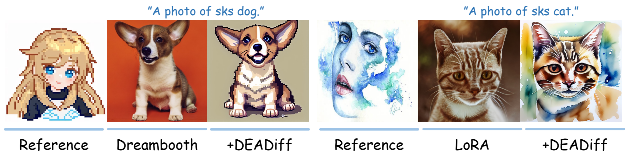

Figure 13. Stylize the Dreambooth/LoRA customized subject.

7.5. Combination with DreamBooth/LoRA

As the original U-Net parameters are frozen, our method is well compatible with DreamBooth&LoRA for extension. Fig. 13 shows an example of using DreamBooth/LoRA to control the subject (the dog and the cat) and DEADiff to control the style.

7.6. More Examples

To show the effectiveness and universality of our method, we present more visualization results in Fig. 14.

[5] https://github.com/lllyasviel/ControlNet-v1-1-nightly#controlnet-11-shuffle

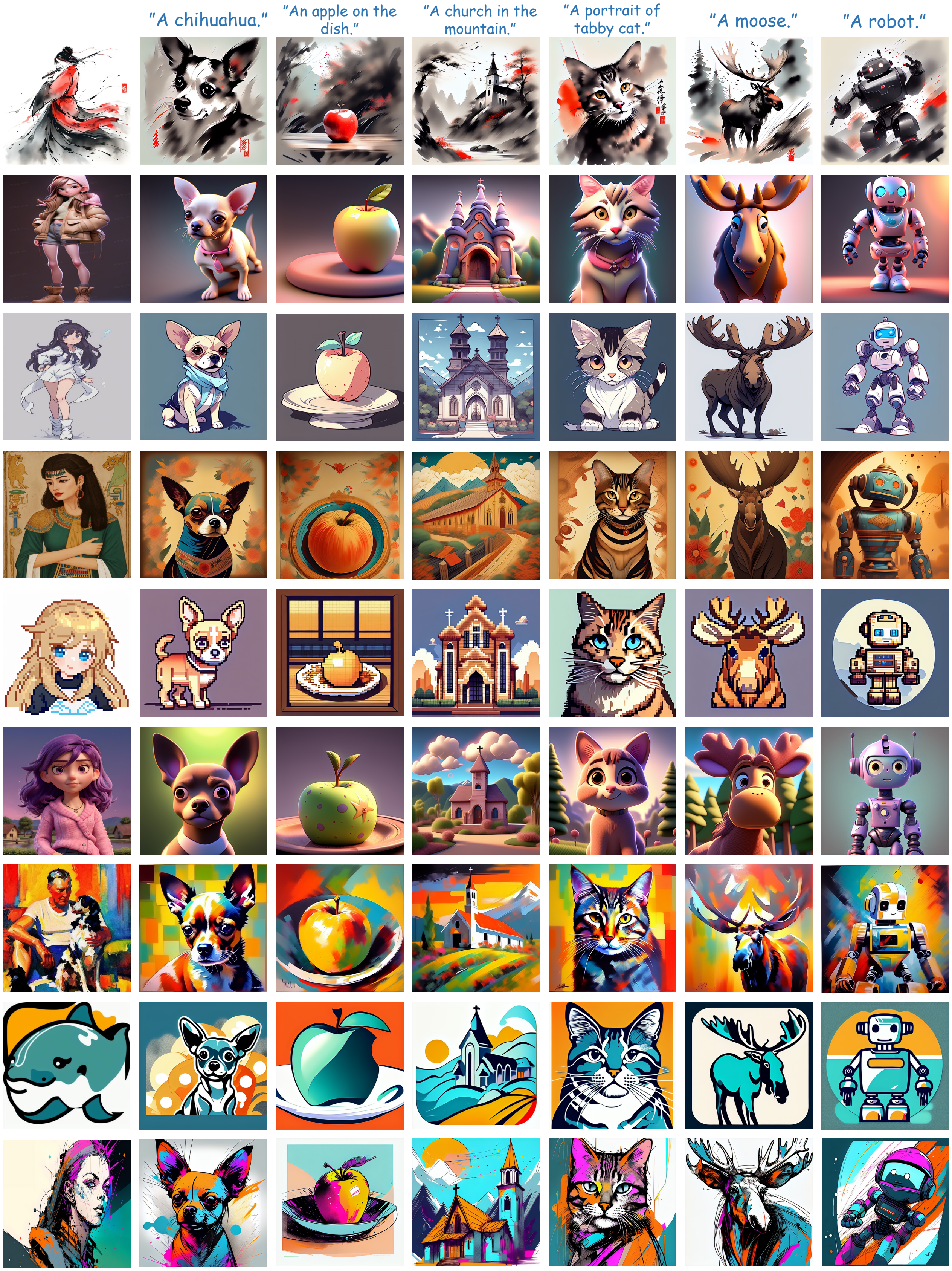

Figure 14. Additional visualization results for *DEADiff*. Our method can synthesize high-quality images that are capable of imitating the reference style and following the instructions of text prompts.