StyleMaster: Stylize Your Video with Artistic Generation and Translation

Supplementary Material

1. Overview

This Supplementary Material is organized into five sections, providing additional details and results to complement the main paper:

- **Section 2:** Provides comprehensive implementation details, including the structure of the base model and the illusion dataset construction.
- **Section 3:** Illustrates complete results of image style transfer.
- **Section 4:** Showcases additional results for stylized video generation.
- **Section 5:** More analysis on illusion dataset.
- **Section 6:** Results of user study.

2. Implementation Details

2.1. Base Model Structure

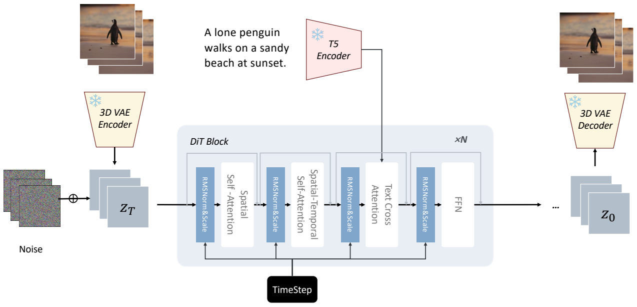


Figure 1. The structure of our base model.

Our model is a DiT-based structure, which consists of a 3D Variational AutoEncoder to convert the video to latent space. Then, the latent feature will pass several DiT blocks [3]. As shown in Fig. 1, each DiT block contains 2D Self-Attention, 3D Self-Attention, Cross-Attention and FFN module. The timestep is embedded as scale and apply RMSNorm to the spatio-temporal tokens before each module.

2.2. Model Illusion Dataset

First, we will make a detailed analysis of the illusion process.

Model Illusion Process. During the generation process using an off-the-shelf T2I model, we can use two different prompts to generate paired image data. To be specific, for a noisy image, we can start a parallel process.

As shown in the Fig. 2, we conduct two transformations on it respectively, in the figure, we use the original image and the vertical flip, which are defined by v_1 and v_2. Then, we use different text prompts to guide the dual-denoising, and we can obtain ϵ_t^1 and ϵ_t^2. Then, we use v_1^{-1} and v_2^{-1} to turn ϵ_t^1 and ϵ_t^2 to the original view, i.e., the view of noisy image. For ϵ_t^1, since v_1 is just itself, so v_1^{-1} will not change ϵ_t^1. For ϵ_t^2, because v_2 is vertical clip, so v_2^{-1} will perform a reversed vertical clip. Then the reversed noise in original view will be added and averaged, to obtain the final ϵ_t.

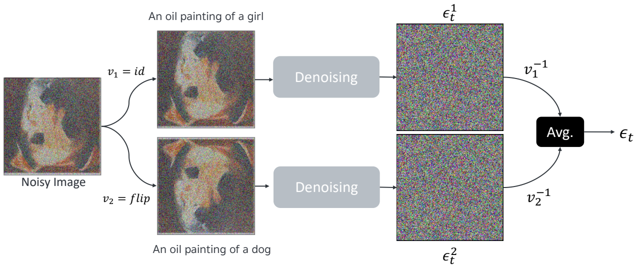


Figure 2. The model illusion process during T2I generation.

During our generation, we set $v_1 = id$ and $v_2 = jigsaw$, the $jigsaw$ means to divide the image into irregular puzzle pieces.

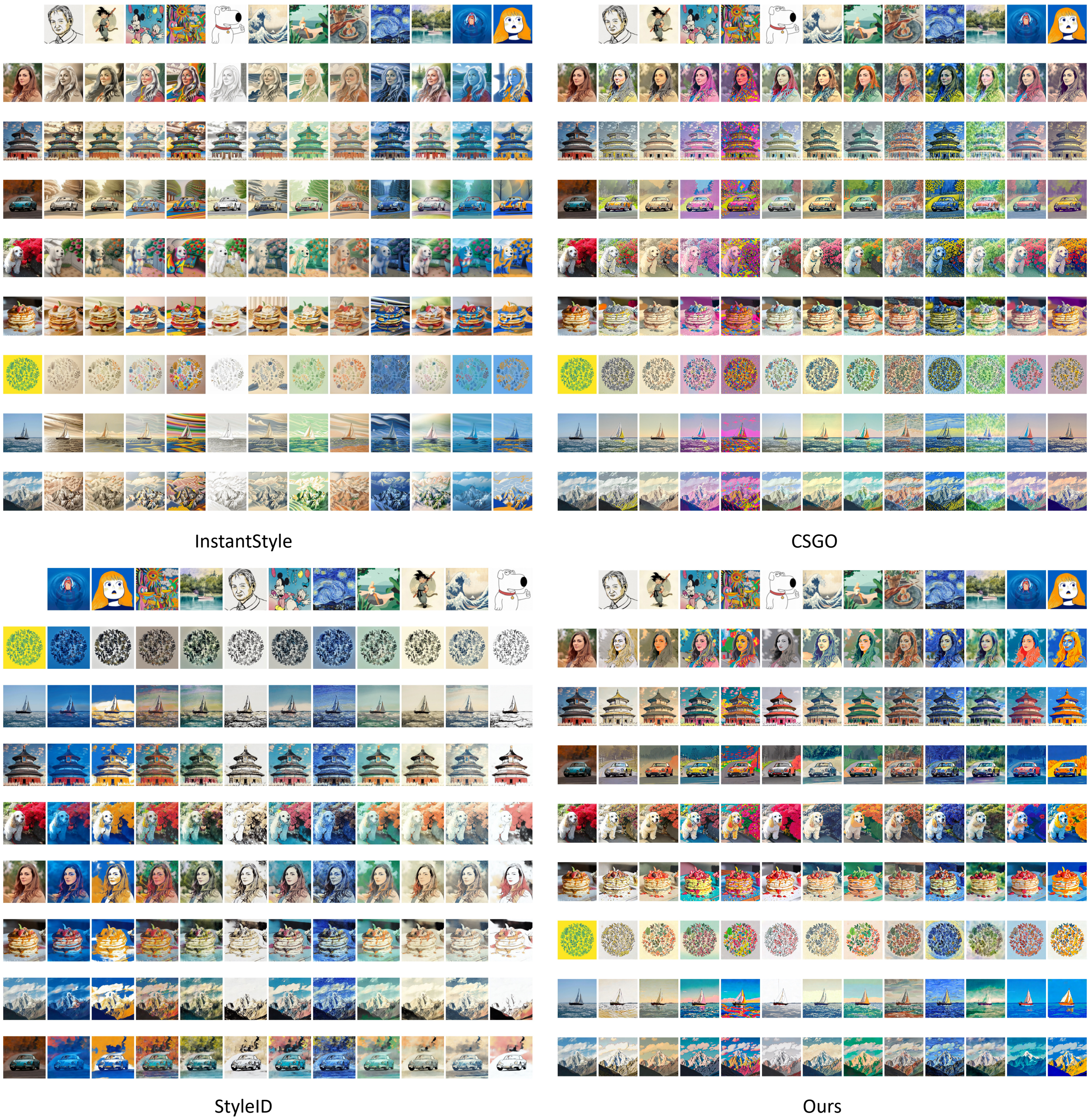
Prompt. For each paired images, we use a pair of prompts to generate them. The prompts are in the form of *a [style] of [object]*. The style and object are from our style list which contains 65 style descriptions collected from the Internet, and the object list consists of 100 common objects. Here we demonstrate 15 samples of each list.

Training. The training loop runs for 100 epochs with a batch size of 8. Two losses are used to supervise this process: triplet loss, which increases the distance between groups, and MSE loss, which reduces the distance within groups. The learning rate is set to 1e-4.

3. Image Style Transfer

Since our method can also be used as an image stylization method, so we compare our method with other image stylization methods, including StyleID [1], InstantStyle [4] and CSGO [6]. We use the default setting in these methods.

Here we illustrate all image style transfer results generated by these three methods with our method in Fig. 3. It is obvious that our method show higher robustness to different styles. For example, StyleID [1] show great ability in transferring the colors, to keep color consistency with reference image. However, it cannot transfer the sematic features of the style.



InstantStyle

CSGO

StyleID

Ours

Figure 3. The image style transfer results generated by four different methods.

4. Stylized Video Generation

More comparison results are shown in Fig. 4. The compared methods are VideoComposer [5] and StyleCrafter [2].

5. Illusion Dataset

We show the filtering process in our dataset construction, as shown in Fig. 5. Negative samples initially lack clear content. By applying CLIP matching score selection, we obtain samples with more appropriate content. Furthermore, training with these higher-quality samples leads to better style representation learning.

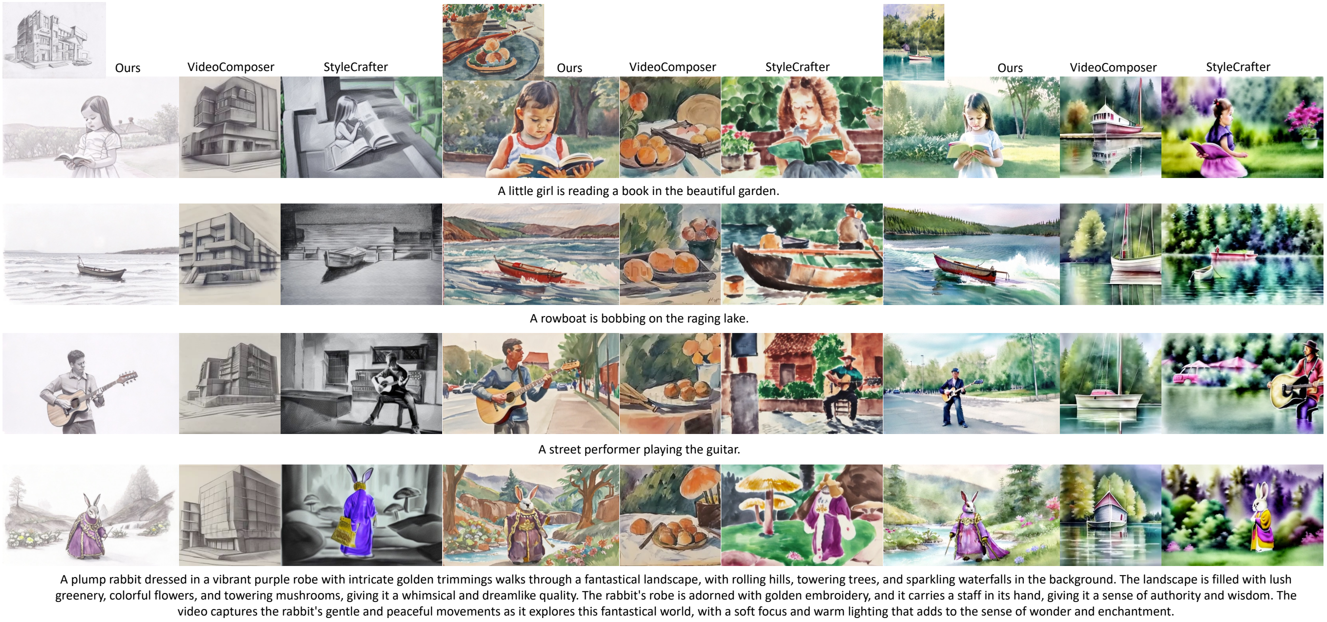

Ours
VideoComposer
StyleCrafter
Ours
VideoComposer
StyleCrafter
Ours
VideoComposer
StyleCrafter
A little girl is reading a book in the beautiful garden.
A rowboat is bobbing on the raging lake.
A street performer playing the guitar.
A plump rabbit dressed in a vibrant purple robe with intricate golden trimmings walks through a fantastical landscape, with rolling hills, towering trees, and sparkling waterfalls in the background. The landscape is filled with lush greenery, colorful flowers, and towering mushrooms, giving it a whimsical and dreamlike quality. The rabbit's robe is adorned with golden embroidery, and it carries a staff in its hand, giving it a sense of authority and wisdom. The video captures the rabbit's gentle and peaceful movements as it explores this fantastical world, with a soft focus and warm lighting that adds to the sense of wonder and enchantment.

Figure 4. **More stylized video generation results.** We compare our method with VideoComposer [5] and StyleCrafter [2].

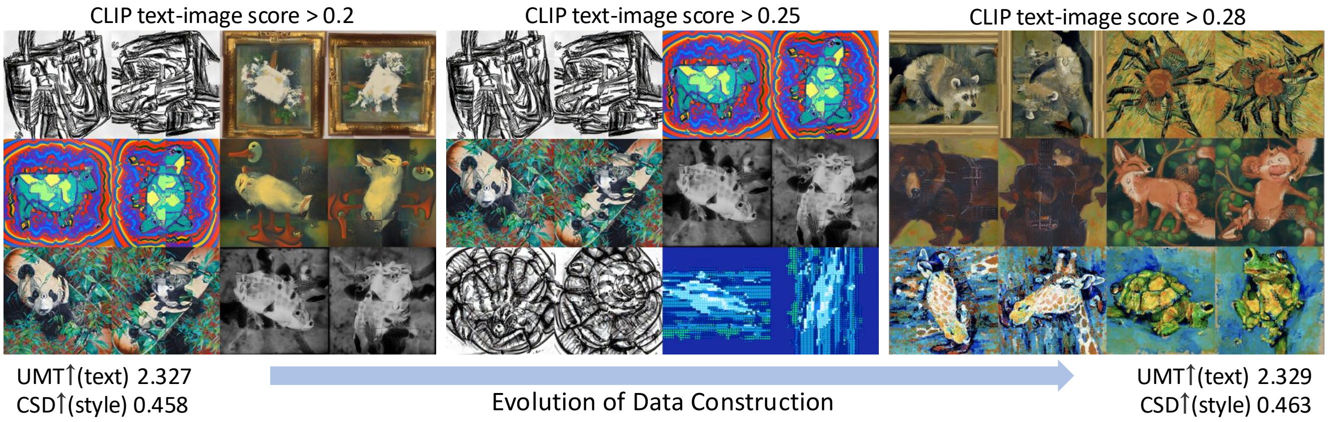

CLIP text-image score > 0.2
CLIP text-image score > 0.25
CLIP text-image score > 0.28
UMT↑(text) 2.327
CSD↑(style) 0.458
Evolution of Data Construction
UMT↑(text) 2.329
CSD↑(style) 0.463

Figure 5. **Data selection of illusion dataset.**

6. User Study

As shown in Table 4, we conduct a user study on stylized generation and style transfer. We invited users to evaluate from three aspects: content/text alignment, style consistency, and temporal behavior. While text alignment is allowed for multiple choices, the other aspects were evaluated using single-choice form. 30 users participated in the study, and we received 912 votes in total. The results show that our method outperforms others in nearly all aspects.

References

[1] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 1

[2] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023. 2, 3

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[4] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 1

[5] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren

object	style
a dog	oil painting
a rabbit	black and white film
a waterfall	cyberpunk picture
a duck	watercolor painting
a teddy bear	vintage photograph
a tudor portrait	3D render
a skull	pencil sketch
houseplants	pop art depiction
flowers	surrealist painting
a landscape	pixel art representation
a boy	comic book illustration
a girl	graffiti street art
an ape	neon-lit cityscape
a parrot	Baroque art piece
a panda	steampunk design

Table 1. Samples from the object list and style list we use for illusion dataset.

A beautiful woman with long, wavy, white hair stands outdoors.
A stunning view of Temple of Heaven in Beijing, showcasing its intricate architecture and against a backdrop of a partly cloudy sky.
A classic car is parked on a winding road with trees in the background.
A fluffy dog sits on a sidewalk next to a vibrant bush of flowers, with rocks in the background.
A stack of pancakes topped with strawberries, nuts, and a drizzle of caramel sauce, served with a side of whipped cream on a plate.
A circular arrangement of various hand-drawn botanical elements, including leaves, branches, and flowers, set against a pure background.
A sailboat glides across sparkling waters under a clear sky, with distant land visible on the horizon.
A majestic snow-capped mountain peak rises against a backdrop of a clear sky dotted with fluffy clouds. The rugged terrain and sharp ridges of the mountain are highlighted by the sunlight, creating a stunning contrast with the shadowed valleys below.

Table 2. The image prompt for image style transfer. Corresponding pictures can refer to Fig. 3.

A little girl is reading a book in the beautiful garden.
A lighthouse is beaming across choppy waters.
A bear is catching fish in a river.
A bouquet of fresh flowers sways gently in the vase with the breeze.
A rowboat is bobbing on the raging lake.
A street performer playing the guitar.
A chef is preparing meals in kitchen.
A student is walking to school with backpack.
A campfire surrounded by tents.
A hot air balloon floating in the sky.
A knight is riding a horse through a field.
A wolf is walking stealthily through the forest.
A river is flowing gently under a bridge.
A lone traveler, dressed in worn clothing and carrying a backpack, walks through a misty forest at sunset. The trees surrounding him are tall and dense, with leaves that are a mix of green and golden hues, reflecting the warm colors of the setting sun. The mist creates a mystical atmosphere, with the air filled with the sweet scent of blooming flowers. The traveler's footsteps are quiet on the damp earth, and the only sounds are the distant chirping of birds and the rustling of leaves.
A group of teddy bears, are holding hands and walking along the street on a rainy day. The bears are dressed in matching raincoats and hats, and they are all smiling and laughing as they stroll along the sidewalk. The rain is coming down gently, and the bears are splashing in the puddles and playing in the rain. The background is a blurred image of the city, with tall buildings and streetlights visible in the distance.
A plump rabbit dressed in a vibrant purple robe with intricate golden trimmings walks through a fantastical landscape, with rolling hills, towering trees, and sparkling waterfalls in the background. The landscape is filled with lush greenery, colorful flowers, and towering mushrooms, giving it a whimsical and dreamlike quality. The rabbit's robe is adorned with golden embroidery, and it carries a staff in its hand, giving it a sense of authority and wisdom. The video captures the rabbit's gentle and peaceful movements as it explores this fantastical world, with a soft focus and warm lighting that adds to the sense of wonder and enchantment.

Table 3. The video prompt for stylized video generation.

Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[6] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 1

	Style Transfer			Stylized Generation		
	InstantStyle +AnyV2V	Domo AI	Ours	Video Composer	Style Crafter	Ours
Content/Text ↑	10.0%	40.0%	**50.0%**	0.017	0.285	**0.911**
Style ↑	5.2%	42.1%	**52.6%**	3.6%	26.8%	**69.6%**
Temporal ↑	2.6%	**50%**	47.3%	14.2%	10.7%	**75.0%**

Table 4. Results of user study.