

# StyleCrafter: Taming Stylized Video Diffusion with Reference-Augmented Adapter Learning

GONGYE LIU, Tsinghua University, China

MENGHAN XIA\*, YONG ZHANG, and HAOXIN CHEN, Tencent AI Lab, China

JINBO XING, The Chinese University of Hong Kong, China

YIBO WANG, Tsinghua University, China

XINTAO WANG and YING SHAN, Tencent AI Lab, China

YUJIU YANG\*, Tsinghua University, China

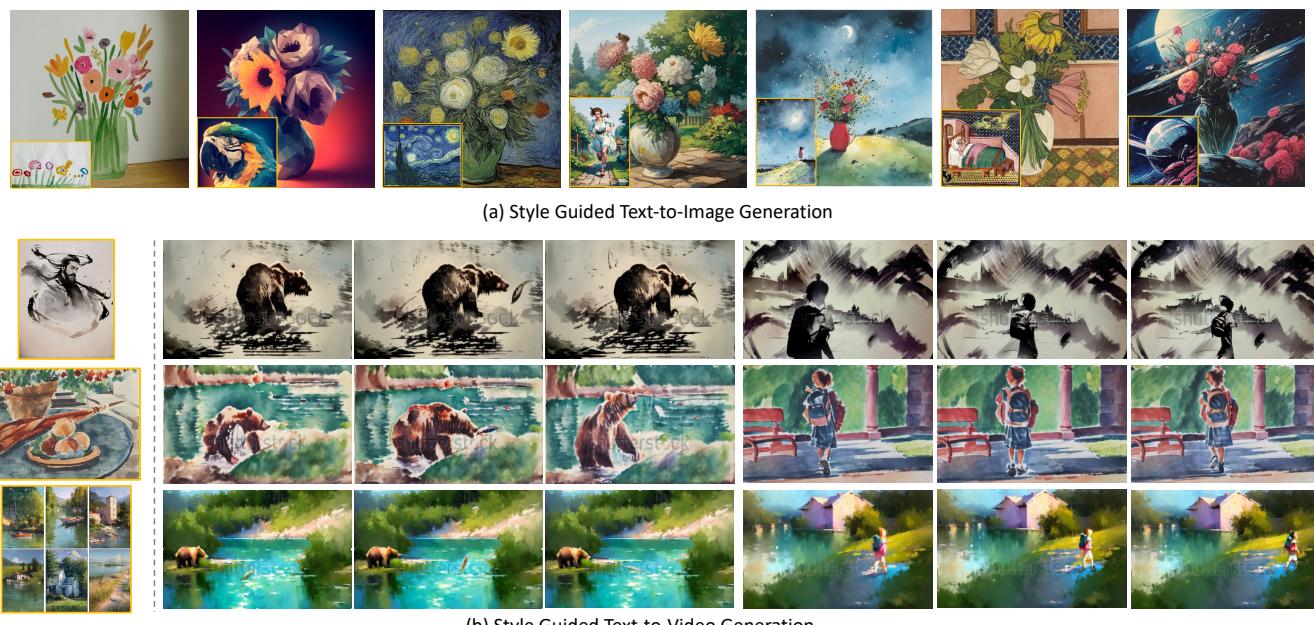


Fig. 1. Stylized Generation Results Produced by StyleCrafter

Text-to-video (T2V) models have shown remarkable capabilities in generating diverse videos. However, they struggle to produce user-desired artistic videos due to (i) text's inherent clumsiness in expressing specific styles and (ii) the generally degraded style fidelity. To address these challenges, we introduce StyleCrafter, a generic method that enhances pre-trained T2V models with a style control adapter, allowing video generation in any style by feeding a reference image. Considering the scarcity of artistic video data, we propose to first train a style control adapter using

\*Corresponding authors

Authors' Contact Information: Gongye Liu, Tsinghua University, Shenzhen, China, lgy22@mails.tsinghua.edu.cn; Menghan Xia; Yong Zhang; Haoxin Chen, Tencent AI Lab, Shenzhen, China, menghanxyz@gmail.com, zhangyong201303@gmail.com, jszxchx@gmail.com; Jinbo Xing, The Chinese University of Hong Kong, Hong Kong, China, jbxing@cse.cuhk.edu.hk; Yibo Wang, Tsinghua University, Shenzhen, China, wyb22@mails.tsinghua.edu.cn; Xintao Wang; Ying Shan, Tencent AI Lab, Shenzhen, China, xintao.alpha@gmail.com, yingshan@tencent.com; Yujiu Yang, Tsinghua University, Shenzhen, China, yang.yujiu@sz.tsinghua.edu.cn.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *ACM Transactions on Graphics*, <https://doi.org/10.1145/3687975>.

style-rich image datasets, then transfer the learned stylization ability to video generation through a tailor-made finetuning paradigm. To promote content-style disentanglement, we employ carefully designed data augmentation strategies to enhance decoupled learning. Additionally, we propose a scale-adaptive fusion module to balance the influences of text-based content features and image-based style features, which helps generalization across various text and style combinations. StyleCrafter efficiently generates high-quality stylized videos that align with the content of the texts and resemble the style of the reference images. Experiments demonstrate that our approach is more flexible and efficient than existing competitors. Project page: <https://gongyeliu.github.io/StyleCrafter.github.io/>

CCS Concepts: • Computing methodologies → Computer Vision.

Additional Key Words and Phrases: Diffusion Model, Stylized Generation, Image/Video Synthesis

ACM Reference Format:

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Ying Shan, and Yujiu Yang. 2024. StyleCrafter: Taming Stylized Video Diffusion with Reference-Augmented Adapter Learning. *ACM Trans. Graph.* 43, 6 (December 2024), 24 pages. <https://doi.org/10.1145/3687975>

## 1 Introduction

The popularity of powerful diffusion models has led to remarkable progress in the field of content generation. For instance, text-to-image (T2I) models are capable of generating diverse and vivid images from text prompts, encompassing various visual concepts. This great success can be attributed not only to the advancement of models but also to the availability of various image data over the Internet. Contrastingly, text-to-video (T2V) models fall short of the data categories especially in styles, since existing videos predominantly feature photorealism. While these strategies, like initializing weights from well-trained T2I models or joint training with image and video datasets, can help mitigate this issue, the generated stylized videos generally suffer from degraded style fidelity. Although significant success has been achieved in style transfer/preservation in T2I generation, the field of stylized video generation remains largely unexplored, and effective solutions are yet to be discovered.

In this paper, we propose StyleCrafter, a generic method that enhances pre-trained T2V models with a style control adapter, enabling text-to-video generation in any desired style by providing a reference image. Anyhow, it is non-trivial to achieve this goal. (i) as a classic problem of style transfer/preservation, the style control adapter requires to extract accurate style concepts from the reference image **in a content-style decoupled manner**. (ii) **the scarcity of open-source stylized videos** challenges the adaptation training of the T2V models.

Considering the scarcity of stylized videos, we propose to first train a style adapter to extract desired style concepts from images over image datasets, and then transfer the learned stylization ability to a T2V model with shared spatial weights through a tailor-made finetuning paradigm. The advantages are twofold: on the one hand, the adapter trained over stylized images can effectively extract the style concept from input images, eliminating the necessity for scarcely available stylized videos. On the other, a finetuning paradigm enables text-to-video models with better adaptation to the style concepts extracted from the previously trained style adapter, while avoiding degradation of temporal quality in video generation.

To effectively capture the style features and promote content-style disentanglement, we adopt the widely used query transformer to extract style concepts from a single image. Particularly, we design a scale-adaptive fusion module to balance the influences of text-based content features and image-based style features, which helps generalization across various text and style combinations. During the training process, we employ carefully designed data augmentation strategies to enhance decoupled learning.

StyleCrafter efficiently generates high-quality stylized videos that align with the content of the texts and resemble the style of the reference images. Comprehensive experiments are conducted to assess our proposed approach, demonstrating that it significantly outperforms existing competitors in both stylized image generation and stylized video generation. Furthermore, ablation studies offer a thorough analysis of the technical decisions made in developing the complete method, which provides valuable insights for the community. Our contributions are summarized as follows:

- We propose the concept of improving stylized generation for pre-trained T2V models by adding a style adapter.

- We explore an efficient network for stylized generation, which facilitates the content-style disentangled generation from text and image inputs. Our method attains notable advantages over existing baselines.
- We propose a training paradigm for generic T2V style adapter without requiring any stylized videos for supervision.

## 2 Related Works

### 2.1 Text to Video Synthesis

Text-to-video synthesis (T2V) is a highly challenging task with significant application value, aiming to generate corresponding videos from text descriptions. Various approaches have been proposed, including autoregressive transformer [Vaswani et al. 2017] models and diffusion models [Ho et al. 2020; Nichol and Dhariwal 2021; Song et al. 2021a,b]. Video Diffusion Model [Ho et al. 2022b] employs a space-time factorized U-Net to execute the diffusion process in pixel space. Imagen Video [Ho et al. 2022a] proposes a cascade diffusion model and v-parameterization to enhance VDM. Another branch of techniques makes good use of pre-trained T2I models and further introduces some temporal blocks for video generation extension. CogVideo [Hong et al. 2022] builds upon CogView2 [Ding et al. 2022] and employs multi-frame-rate hierarchical training strategy to transition from T2I to T2V. Similarly, Make-a-video [Singer et al. 2022], MagicVideo [Zhou et al. 2022] and LVDM [He et al. 2022] inherit pretrained T2I diffusion models and extend them to T2V generation by incorporating temporal attention modules.

### 2.2 Stylized Image Generation

Stylized image generation aims to create images that exhibit a specific style. Decoupling style and content is a classic challenge [Tenenbaum and Freeman 2000]. Early research primarily concentrated on image style transfer, a technique that involves the transfer of one image's style onto the content of another, requiring a source image to provide content. Traditional style transfer methods [Hertzmann et al. 2001; Wang et al. 2004; Zhang et al. 2013] employ low-level, hand-crafted features to align patches between content images and style images. Since Gatys et al. [Gatys et al. 2016] discovered that the feature maps in CNNs capture style patterns effectively, a number of studies [An et al. 2021; Deng et al. 2022; Huang and Belongie 2017; Li et al. 2017; Liu et al. 2021; Texler et al. 2020a; Zhang et al. 2022] have been denoted to utilize neural networks to achieve arbitrary style transfer. A common practice involves utilizing a pretrained VGG network [Simonyan and Zisserman 2014] to extract style information or compute Gram matrix loss [Gatys et al. 2016] to enable self-supervised learning of visual styles.

As the field of generation models progressed, researchers began exploring stylized image generation for T2I models. Although T2I models can generate various artistic images from corresponding text prompts, words are often limited to accurately convey the stylistic elements in artistic works. Consequently, recent works have shifted towards example-guided artistic image generation. Several studies [Hu et al. 2022; Kumari et al. 2023; Ruiz et al. 2023; Shi et al. 2023] developed various optimization techniques on a small collection of input images that share a common style concept. Inspired by Textural Inversion (TI) [Gal et al. 2022], some methods [Ahn et al.

2023; Sohn et al. 2023; Zhang et al. 2023] propose to optimize a specific textual embedding to represent a certain style. Similarly to our work, IP-Adapter [Ye et al. 2023] trains an image adapter based on pretrained Stable Diffusion to adapt T2I models to image conditions. Although IP-Adapter can produce similar image variants, it fails to decouple style concepts from input images or generate images with other content through text conditions.

### 2.3 Stylized Video Generation

Building upon the foundation of stylized image generation, researchers have extended the concept to video style transfer and stylized video generation. Due to the scarcity of large-scale stylized video data, a common approach for video stylization involves applying image stylization techniques on a frame-by-frame basis. Before the advent of ML, researchers have explored methods for rendering specific artistic styles such as video watercolorization [Bousseau et al. 2007]. Early deep learning methods of video style transfer [Chen et al. 2017; Deng et al. 2021; Gao et al. 2020; Jamriška et al. 2019; Ruder et al. 2016; Texler et al. 2020b] apply style transfer in video sequences, generating stable stylized video sequences through the use of optical flow constraints. Additionally, Some video editing methods [Geyer et al. 2024; Huang et al. 2023b; Khachatryan et al. 2023; Qi et al. 2023; Wu et al. 2023; Yang et al. 2023, 2024] based on pretrained T2I models also support text-guided video style transfer. Although these methods effectively improve temporal consistency, they often fail to handle frames with a large action span. Reliance on a source video also undermines flexibility. Similarly, certain image-to-video(I2V) methods [Blattmann et al. 2023; Xing et al. 2024, 2023] demonstrate capabilities in stylized video generation, particularly in the anime domain. However, I2V models still face challenges when tasked with interpreting and animating highly artistic images, producing frames that veer towards realism, since real-world videos dominated its training data.

VideoComposer [Wang et al. 2024] focuses on controllable video generation, allowing multiple conditional input to govern the video generation, including structure, motion, style, etc. Although VideoComposer enables multiple controls including style, they fail to decouple style concepts, leading to limited visual quality and motion naturalness. AnimateDiff [Guo et al. 2024] employs a T2I model as a base generator and adds a motion module to learn motion dynamics, which enables extending the success of personalized T2I models(e.g., LoRA [Hu et al. 2022], Dreambooth [Ruiz et al. 2023]) to video animation. However, the dependence on a personalized model restricts its ability to generate videos with arbitrary styles. Another associated research is Text2Cinemagraph [Mahapatra et al. 2023], which utilizes pretrained text-to-image models to pioneer text-guided artistic cinemagraph creation. This approach surpasses some existing text-to-video models like VideoCrafter [Chen et al. 2023a] in generating plausible motion in artistic scenes. Nevertheless, its main limitation lies in its confined applicability, primarily to landscapes, and its tendency to generate scanty motion patterns solely for fluid elements.

## 3 Method

We propose a method to equip pre-trained Text-to-Video (T2V) models with a style adapter, allowing for the generation of stylized

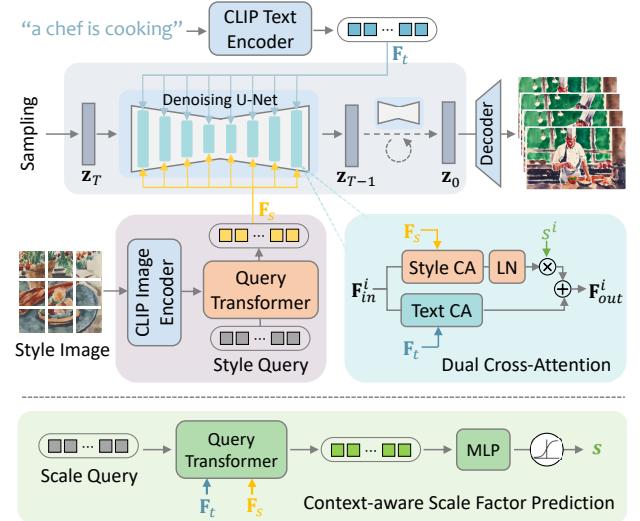


Fig. 2. Overview of our proposed style adapter. It consists of three components, i.e. style feature extractor, dual cross-attention module, and context-aware scale factor predictor.

videos based on both a text prompt and a style reference image. The overview is illustrated in Figure 2. In this framework, the textual description dictates the video content, while the style image governs the visual style, ensuring a disentangled control over the video generation process. Given the limited availability of stylized videos, we employ a two-stage training strategy. Initially, we utilize an image dataset abundant in artistic styles to learn reference-based style modulation. Subsequently, adaptation finetuning on a mixed dataset of style images and realistic videos is conducted to improve the temporal quality of the generated videos.

### 3.1 Reference-Based Style Modulation

Our style adapter serves to extract style features from the input reference image and infuse them into the backbone features of the denoising U-Net. As mainstream T2V models [Chen et al. 2023a, 2024; Wang et al. 2023c,a] are generally initialized from open-source T2I Models and trained with image and video datasets in a joint strategy, they support not only text-to-video generation but also retain the capacity for text-to-image generation. To overcome the scarcity of stylized videos, we propose to train the style adapter based on a pre-trained T2V model (i.e. VideoCrafter [Chen et al. 2023a]) for stylized image generation under the supervision of stylistic images.

*Content-Style Decoupled Data Augmentation.* We use the stylistic images from two publicly available datasets, i.e. WikiArt [Phillips and Mackintosh 2011] and a subset of Laion-Aesthetics [Schuhmann et al. 2022] (aesthetics score above 6.5). In the original image-caption pairs, we observe that the captions generally contain both content and style descriptions, and some of them do not match the image content well. To promote the content-style decoupling, we use BLIP-2 [Li et al. 2023a] to regenerate captions for the images and remove certain forms of style description (e.g., *a painting of*) with regular expressions. In addition, as an image contains both style and content information, it is necessary to construct a decoupling supervision strategy to guarantee the extracted style feature free of content features. Although a stylistic image may contain different

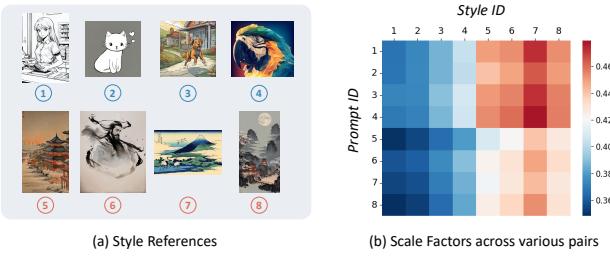


Fig. 3. Illustration of content-style fusion scale factors across multiple input pairs. Four short prompts (less than 5 words) with prompt id  $\in [1, 4]$  and four long prompts (more than 8 words) with prompt id  $\in [5, 8]$  are randomly selected. Results indicate that shorter prompts and images with richer style-semantics tend to have relatively higher scale factors.

local style patterns [Chen et al. 2023b; Huo et al. 2021; Park and Lee 2019], we regard that a large crop of an image (e.g. 50% of the image) still preserves a similar style representation with the full image. Based on this insight, we process each stylistic image to obtain the target image and style image through different strategies: for target image, we scale the shorter side of the image to 512 and then crop the target content from the central area; for style image, we scale the shorter side of the image to 800 and randomly crop a local patch with  $512 \times 512$ . This approach reduces the overlap between the style reference and generation target, while still preserving the global style semantics complete and consistent.

*Style Embedding Extraction.* CLIP [Radford et al. 2021] has demonstrated remarkable capability in extracting visual features from open-domain images. To capitalize on this advantage, we employ a pre-trained CLIP image encoder as a feature extractor. Specifically, we utilize both the global semantic token and the full 256 local tokens (i.e., from the final layer of the Transformer) since our desired style embedding should not only serve as an accurate style trigger for the T2V model, but also provide auxiliary feature references. As image tokens encompass both style and content information, we further employ a trainable Query Transformer (Q-Former) [Li et al. 2023a] to extract style embedding  $F_s$ . We create  $N$  learnable style query embeddings as input for the Q-Former, which interact with image features through self-attention layers. Note that this is a commonly adopted architecture for visual condition extraction [Li et al. 2023a; Shi et al. 2023; Xing et al. 2023; Ye et al. 2023]. But it is the style-content fusion mechanism that makes our proposed design novel and insightful for style modulation, as detailed below.

*Adaptive Style-Content Fusion.* With the extracted style embedding, there are two ways to combine the style and text conditions, including (i) *attach-to-text* [Huang et al. 2023a; Li et al. 2023b; Ramesh et al. 2022]: attach the style embedding to the text embedding and then interact with the backbone feature via the originally text-based cross-attention as a whole; (ii) *dual cross-attention* [Wei et al. 2023; Ye et al. 2023]: adding a new cross-attention module for the style embedding and then fuse the text-conditioned feature and style-conditioned feature. According to our experiment (see Sec. 4.4), solution (ii) surpasses solution (i) in disentangling the roles of text and style conditions, therefore we have adopted it as our final solution. The formula can be written as:

$$\mathbf{F}_{out}^i = \text{TCA}(\mathbf{F}_{in}^i, \mathbf{F}_t) + s^i * \text{LN}(\text{SCA}(\mathbf{F}_{in}^i, \mathbf{F}_s)), \quad (1)$$

where  $\mathbf{F}_{in}^i$  denotes the backbone feature of layer  $i$ , LN denotes layer normalization, and TCA and SCA denote text-based cross attention and style-based cross attention respectively.  $s^i$  is a scale factor learned by a context-aware scale factor prediction network, to balance the magnitudes of text-based feature and style-based feature. The motivation is that different stylistic genres may have different emphasis on content expression. For example, the abstract styles tend to diminish the concreteness of the content, while realism styles tend to highlight the accuracy and specificity of the content. So, we propose a context-aware scale factor prediction network to predict fusion scale factors according to the input contexts. Specifically, we create a learnable factor query, it interacts with textual features  $\mathbf{F}_t$  and style features  $\mathbf{F}_s$  to generate scale features via a Q-Former and then project it into layer-wise scale factors  $\mathbf{s} \in \mathbb{R}^{16}$ . Figure 3 illustrates the learned scale factors across multiple contexts. It shows that the adaptive scale factors have a strong correlation with style genres while also depending on the text prompts. Style references with rich style-semantics (i.e., ukiyo-e style) typically yield higher scale factors to emphasize style; while complex prompts tend to produce lower scale factors to enhance content control. This is consistent with our hypothesis to motivate our design.

### 3.2 Temporal Adaptation to Stylized Features

Given a pre-trained T2V model, the style adapter trained on image dataset works well for stylized image generation. However, it still struggles to generate satisfactory stylized videos, which is vulnerable to temporal jittering and visual artifacts. The possible causes are that the cross-frame operations, i.e. temporal self-attention, do not involve in the process of stylized image generation, and thus induce incompatible issues. So, it is necessary to finetune the temporal self-attention with the style adapter incorporated. Following the practice of T2V image and video joint training, the finetuning is performed on the mixed datasets of stylistic images and photorealistic videos. This is an adaptation training of temporal blocks while the other modules remain frozen, and the model converges efficiently.

*Classifier-Free Guidance for Multiple Conditions.* Unlike T2I models, video models exhibit a higher sensitivity to style guidance due to their limited stylized generation capabilities. Using a unified  $\lambda$  for both style and context guidance may lead to undesirable generation results. Regarding this, we adopt a more flexible mechanism for multiple conditions classifier-free guidance. Building upon the vanilla text-guided classifier-free guidance, which controls context alignment by contrasting textual-conditioned distribution  $\epsilon(z_t, c_t)$  with unconditional distribution  $\epsilon(z_t, \emptyset)$ , we introduce the style guidance with  $\lambda_s$  by emphasizing the difference between the text-style-guided distribution  $\epsilon(z_t, c_t, c_s)$  and the text-guided distribution  $\epsilon(z_t, c_t)$ . The complete formulation is as below:

$$\begin{aligned} \hat{\epsilon}(z_t, c_t, c_s) &= \epsilon(z_t, \emptyset) + \lambda_s (\epsilon(z_t, c_t, c_s) - \epsilon(z_t, c_t)) \\ &\quad + \lambda_t (\epsilon(z_t, c_t) - \epsilon(z_t, \emptyset)), \end{aligned} \quad (2)$$

where  $c_t$  and  $c_s$  denote textual and style condition respectively.  $\emptyset$  denotes using no text or style conditions. In our experiment, we follow the recommended configuration of text guidance in VideoCrafter [Chen et al. 2023a], setting  $\lambda_t = 15.0$ , while the style

guidance is configured with  $\lambda_s = 7.5$  empirically. Similarly, we set  $\lambda_t = 7.5$  and  $\lambda_s = 5.0$  for style-guided image generation.

## 4 Experimental Results

### 4.1 Experimental settings

*Implementation Details.* We adopt the VideoCrafter [Chen et al. 2023a] as our base T2V model, which shares the same spatial weights with Stable Diffusion 2.1. We first train the style modulation on image dataset, i.e. WikiArt [Phillips and Mackintosh 2011] and Laion-Aesthetics-6.5+ [Schuhmann et al. 2022] for 40k steps with a batch size of 32 per GPU. In the second stage, we froze the style modulation part and only train temporal blocks of VideoCrafter, we jointly train image datasets and video datasets(subset of WebVid-10M [Bain et al. 2021]) for 20k steps with a batch size of 1 on video data and 16 on image data, sampling image batches with a ratio of 20%. The training process is performed on 8 A100 GPUs and can be completed within 3 days. Furthermore, to ensure a fair comparison with some SDXL-based models [Hertz et al. 2023; Ye et al. 2023] on stylized image generation, we also trained the first stage of StyleCrafter on SDXL [Podell et al. 2023a].

*Testing Datasets.* To evaluate the effectiveness and generalizability of our method, we construct testsets comprising content prompts and style references. For content prompts, we use GPT-4 [OpenAI 2023] to generate recognizable textual descriptions from four meta-categories (human, animal, object, and landscape). We manually filter out low-quality prompts, retaining 20 image prompts and 12 video prompts. For style references, we collect 20 stylized images and 8 sets of style images with multi-reference (each contains 5 to 7 images in similar styles) from the Internet. In total, the test set contains 400 pairs for stylized image generation, and 300 pairs for stylized video generation (240 single-reference pairs and 60 multi-reference pairs). Details are available in the supplementary materials.

*Evaluation Metrics.* Following previous practice [Sohn et al. 2023; Wang et al. 2023b; Zhang et al. 2023], we employ CLIP-based [Radford et al. 2021] scores and DINO-based [Caron et al. 2021] scores to measure the text alignment and style conformity. Following Eval-Crafter [Liu et al. 2023], we measure the temporal consistency of video generation by (i) calculating clip scores between contiguous frames and (ii) calculating the warping error on every two frames with estimated optical flow. Note that these metrics are not perfect. For example, one can easily achieve a close-to-1 style score by entirely replicating the style reference. Similarly, stylized results may yield inferior text scores compared to realistic results, even though both accurately represent the content descriptions. We recommend a comprehensive consideration of both CLIP-based text scores and style scores, rather than relying solely on a single metric.

*User Preference Study.* In addition to quantitative analysis, we conducted a user study to make comparisons among our method, VideoCrafter, Gen-2, and AnimateDiff in the context of single-reference and multi-reference stylized video generation. Users are instructed to select their preferred option based on style conformity, temporal quality, and all options fulfill text alignment for each comparison pair. We randomly chose 15 single-reference pairs and 10

multi-reference pairs, collecting 1125 votes from 15 users. Further details can be found in the supplementary materials.

### 4.2 Style-Guided Text-to-Image Generation

As mentioned in Sec. 3.1 and Sec. 4.1, our proposed method also supports to generate stylized images (using model before temporal finetuning). We are interested to evaluate our method against state-of-the-art style-guided T2I synthesis methods, which are better-established than video counterparts. The competitors include optimization-based methods like DreamBooth [Ruiz et al. 2023], inversion-based methods such as InST [Zhang et al. 2023] and Style-Aligned [Hertz et al. 2023], and adapter-based methods like IP-Adapter-Plus [Ye et al. 2023]. Besides, we consider two unique competitors: SD\* [Rombach et al. 2022] and SDXL\* [Podell et al. 2023b] (text-to-image models equipped with GPT-4V [OpenAI 2023], where GPT-4V generates textual descriptions about the reference’s style and merges them with content prompts as input for models). This comparison aims to validate the advantages of employing image conditions to enhance stylized generation instead of relying solely on text conditions. Implementation details of competitors are available in supplementary materials.

The quantitative comparison is tabulated in Table 1. Results reveal that Dreambooth [Ruiz et al. 2023] and InST [Zhang et al. 2023] struggle to accurately capture the style from various style references and exhibit low style conformity. SD\* [Rombach et al. 2022] and SDXL [Podell et al. 2023b] demonstrate good stylistic ability but still fail to reproduce the style of the reference image, possibly because of the text’s inherent clumsiness in expressing specific styles despite utilizing the powerful GPT4V for visual style understanding. IP-Adapter [Ye et al. 2023] and style-aligned [Hertz et al. 2023] generate aesthetically pleasing images, while their style-content decoupled learning is not perfect and exhibits limited control over content. In contrast, our method efficiently generates high-quality stylized images that align with the content of the texts and resemble the style of the reference image. Our method demonstrates stable stylized generation capabilities when dealing with various types of prompts.

### 4.3 Style-Guided Text-to-Video Generation

Existing approaches for style-guided video generation can be divided into two categories: one is the single-reference based methods that are usually tuning-free, e.g. VideoComposer [Wang et al. 2024]; the other is the multi-reference based methods that generally requires multiple references for fine-tuning, e.g. AnimateDiff [Guo et al. 2024]. We make comparisons with these methods, respectively.

*Single-Reference based Guidance.* VideoComposer [Wang et al. 2024] is a controllable video generation model that allows multiple conditional inputs including style reference image. It is a natural competitor of our method. Besides, we construct two additional comparative methods, i.e. VideoCrafter\* and Gen2\*, which extend VideoCrafter [Chen et al. 2023a] and Gen2 [Gen-2 2023], the state-of-the-art T2V models in open-source and close-source channels respectively, to make use of style reference images by utilizing GPT-4V [OpenAI 2023] to generate style prompts from them. The quantitative comparison is tabulated in Table 2. Several typical visual examples are illustrated in Figure 5.

Table 1. Quantitative comparison on single-reference style-guided T2I generation. We conduct evaluation on a test set of 400 pairs. **Bold**: Best.

Method	Stable Diffusion 2.1 based				SDXL based			
	Dreambooth	InST	SD*	Ours	IP-Adapter-Plus	Style-Aligned	SDXL*	Ours(SDXL)
CLIP-Text ↑	<b>0.3047</b>	0.3004	0.2766	0.3028	0.2768	0.2254	0.2835	<b>0.2918</b>
CLIP-Style ↑	0.3459	0.3708	0.4183	<b>0.4836</b>	0.5182	0.5515	0.4348	<b>0.5615</b>
DINO-Style ↑	0.2278	0.2587	0.2890	<b>0.3652</b>	0.4367	0.4395	0.2912	<b>0.4514</b>



Fig. 4. Visual comparison on style-guided T2I generation. Blue: methods based on SD 2.1. Green: based on SDXL. Prompt: A rabbit nibbling on a carrot.

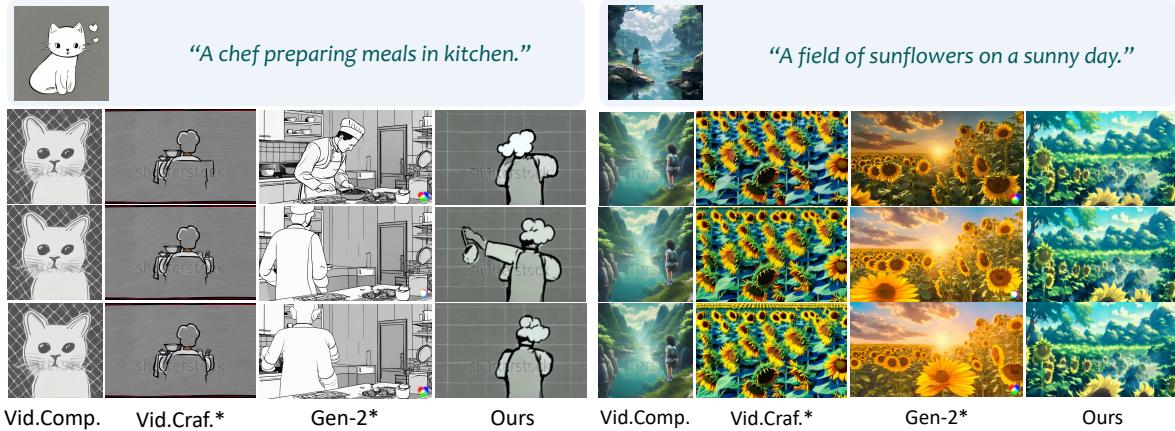


Fig. 5. Visual comparison of single-reference guided T2V generation. Vid.Comp.: VideoComposer, Vid.Craf.: VideoCrafter

Table 2. Quantitative comparison of style-guided T2V generation. Top 3 rows: single-reference based guidance. Bottom 3 rows: multi-reference based guidance. S-R: Single-Reference, M-R: Multi-Reference, W.E.: Warping Errors.

Methods	CLIP-Text ↑	CLIP-Style ↑	Temporal Consistency	
			CLIP-Temp ↑	W.E. ( $\times 10^{-3}$ ) ↓
VideoComposer	0.0468	<b>0.7306</b>	0.9853	<b>9.903</b>
VideoCrafter*	0.2209	0.3124	0.9757	61.41
Ours	<b>0.2726</b>	0.4531	<b>0.9892</b>	18.73
AnimateDiff	<b>0.2867</b>	0.3528	0.8903	37.17
Ours(S-R)	0.2661	0.4803	0.9851	14.13
Ours(M-R)	0.2634	<b>0.4887</b>	<b>0.9852</b>	<b>9.396</b>

We can observe that: (i) VideoComposer tends to copy content from style references and struggles to generate text-aligned content,

Table 3. User study statistics of the selection rate for text alignment(Text), and preference rate for style conformity(Style) and temporal quality(Temporal). Top 3 rows: single-reference based guidance. Bottom 2 rows: multi-reference based guidance.

Methods	Text ↑	Style ↑	Temporal ↑
VideoCrafter*	0.391	8.0%	4.4%
Gen-2*	0.747	23.1%	51.1%
Ours	0.844	68.9%	44.4%
AnimateDiff	0.647	10.0%	19.3%
Ours(M-R)	0.907	90.0%	80.7%

which is possibly because of the invalid decoupling learning. Consequently, its results exhibit abnormally high style conformity and

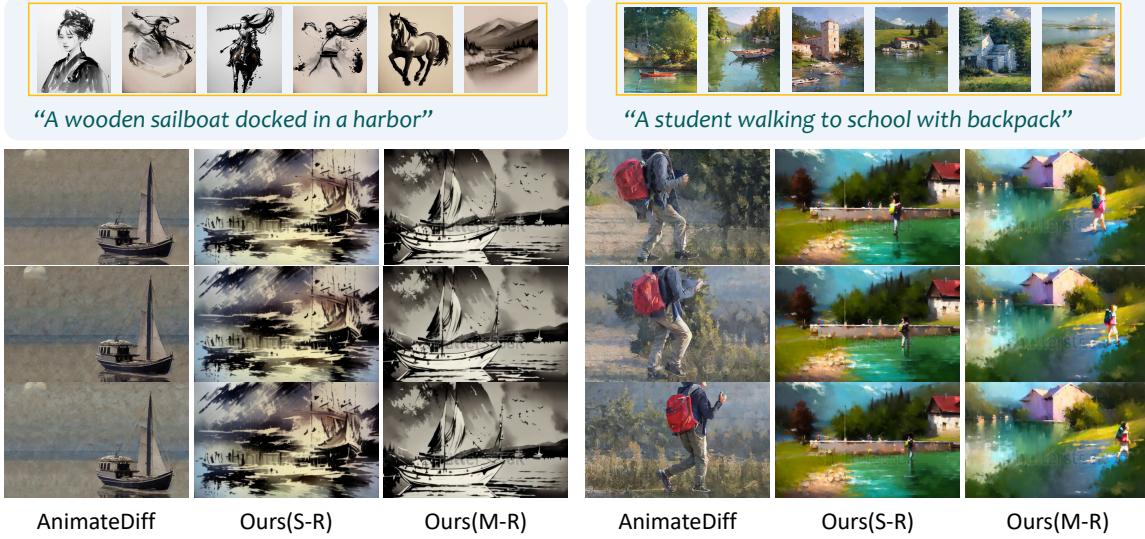


Fig. 6. Qualitative comparison of multi-reference style-guided T2V generation. S-R: Single-Reference, M-R: Multi-Reference

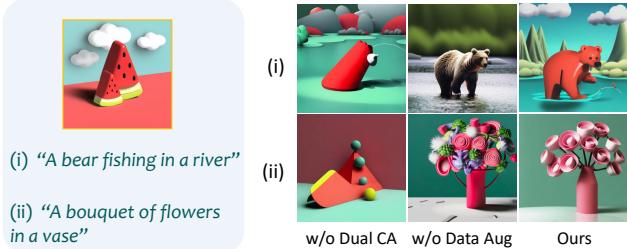


Fig. 7. Effects of dual cross-attention and data augmentation.

very low text alignment. In addition, VideoComposer often generates static videos, thus having the lowest warping errors, but this does not mean that their results perform best in temporal quality. (ii) VideoCrafter\* exhibits limited stylized generation capabilities, producing videos with diminished style and disjointed movements. Gen-2\* demonstrates superior stylized generation capabilities. However, Gen-2 is still limited by the inadequate representation of style in textual descriptions, and is more prone to sudden changes in color and luminance. (iii) In comparison, our method captures styles more effectively and reproduces them in the generated results.

*Multi-Reference based Guidance.* AnimateDiff [Guo et al. 2024] denotes a paradigm to turn personalized SD (i.e., SD fine-tuned on specific-domain images via LoRA [Hu et al. 2022] or Dreambooth [Ruiz et al. 2023]) for video generation, namely combined with pre-trained temporal blocks of T2V models. It can generate very impressive results if the personalized SD is carefully prepared, however, we find it struggle to achieve as satisfactory results if only a handful of style reference images are available for training. We conduct an evaluation on 60 text-style pairs with multi-references, as presented in Sec. 4.1. We train Dreambooth [Ruiz et al. 2023] models for each style and incorporate them into AnimateDiff based on their released codebase. Thanks to the flexibility of Q-Former, our method also supports multiple reference images in a tuning-free fashion, i.e. computing the image embeddings of each reference

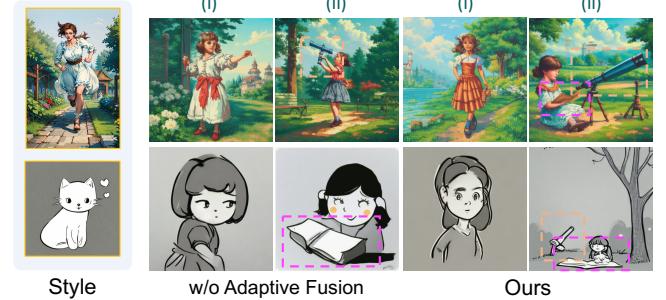


Fig. 8. Effect of adaptive content-style fusion. It shows superiority in generalization to extreme cases, e.g. long text description. Two text prompts are used: (i) A little girl; (ii) A little girl *reading a book* in the park, with a telescope nearby pointed at the sky.

image and concatenating all embeddings as input to the Q-Former. Results are compared in Table 3 and Figure 6 respectively.

According to the results, AnimateDiff struggles to achieve high-fidelity stylistic appearance while tends to generate close-to-realism results despite the style references are typical artistic styles. In addition, it is vulnerable to temporal artifacts. As the trained personalized-SD can generate decent stylistic images (provided in the supplementary materials), we conjecture that the performance degradation is caused by the incompatibility from the pre-trained temporal blocks and independently trained personalized-SD models, which not only interrupts temporal consistency but also weakens the stylistic effect. In contrast, our method can generate temporal consistent videos with high style conformity to the reference images and accurate content alignment with the text prompts. Furthermore, using multiple references can further promote the performance, which offers additional advantages in practical applications.

#### 4.4 Ablation Study

*Data Augmentation.* We first study the effectiveness of content-style decoupled data augmentation. As depicted in Table 4, training

Table 4. Ablation studies on style modulation designs. The performance is evaluated based on the style-guided T2I generation.

Methods	CLIP-Text ↑	CLIP-Style ↑
Ours	0.3028	0.4836
w/o Data Augmentation	0.3173	0.4005
w/o Dual Cross Attention	0.0983	0.7332
w/o Adaptive Fusion	0.2807	0.4925

Table 5. Ablation study on our two-stage training scheme.

Methods	CLIP-Text ↑	CLIP-Style ↑	Temporal Consistency	
			CLIP-Temp ↑	W.E.( $\times 10^{-3}$ ) ↓
w/o Temporal Adaption	0.2691	0.3923	0.9612	47.88
Joint Training	0.3138	0.2226	0.9741	24.74
Two-Stage(ours)	<b>0.2726</b>	<b>0.4531</b>	<b>0.9892</b>	<b>18.73</b>

with the original image-caption pairs restricts the model's ability to extract style representations, leading to lower style conformity. For example, as shown in Figure 7, method without data augmentation fails to capture the "3D render" style from the reference.

**Dual Cross-Attention.** As discussed in Sec. 3.1, we make a comparison between *attach-to-text* and **dual cross-attention** to study their effects. Results are presented in Table 4 and Figure 7, revealing that *attach-to-text* tends to directly fuse the content from the reference image and the text prompts rather than combining the text-based content and image-based style. This indicates the effectiveness of **dual cross-attention** in facilitating content-style decoupling.

**Adaptive Style-Content Fusion.** As previously discussed in Figure 3, our proposed adaptive style-content fusion module demonstrates effectiveness in adaptively processing various conditional contexts. It benefits the generalization ability of model to deal with diverse combinations of content prompt and style image. Figure 8 reveals that although the baseline can handle easy prompt inputs like "A little girl", it struggles to accurately generate all objects described in longer prompts. In contrast, the adaptive fusion module can achieve decent text alignment for long text descriptions thanks to its flexibility to adaptive balance between content and style.

**Two-Stage Training Scheme.** Our proposed training scheme consists of two stages, i.e., style adapter training and temporal adaption. To show its necessity, we build two baselines: (i) *w/o Temporal Adaption*: that we train a style adapter on image data and apply it directly to stylized video generation without finetuning; (ii) *joint training*: that we conduct style adapter training and temporal blocks finetuning on image-video dataset simultaneously. As depicted in Table 5, baseline (i) exhibits inferior temporal consistency when applied directly to video, and undermines the content alignment and style conformity. As for baseline (ii), the learning of style embedding extraction seems to be interfered by the joint finetuning of temporal blocks, which impedes it to generate desirable stylized videos.

## 5 Conclusion and Limitations

We have presented StyleCrafter, a generic method enabling pre-trained T2V model for video generation in any style by providing a reference image. To achieve this, we made exploration in three aspects, including the architecture of style adapter, the content and style feature fusion mechanism, and some tailor-made strategies for

data augmentation and training stylized video generation without stylistic video data. All of these components allow our method to generate high quality stylized videos that align with text prompts and conform to style references. Extensive experiments have evidenced the effectiveness of our proposed designs and comparisons with existing competitors demonstrate the superiority of our method in visual quality and efficiency. Anyway, our method also has certain limitations, e.g., unable to generate desirable results when the reference image can not represent the target style sufficiently or the presented style is extremely unseen. Further explorations are demanded to address those issues.

## ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61991451) and the Shenzhen Science and Technology Program (JSGG20220831093004008).

## References

- Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeam Hong. 2023. DreamStyler: Paint by Style Inversion with Text-to-Image Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 862–871.
- Max Bain, Arsha Nagrani, Gùl Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1728–1738.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- Adrien Bousseau, Fabrice Neyret, Joëlle Thollot, and David Salesin. 2007. Video watercolorization using bidirectional texture advection. *ACM Transactions on Graphics (ToG)* 26, 3 (2007), 104–es.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.
- Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*. 1105–1114.
- Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. 2023a. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *preprint arXiv:2310.19512* (2023).
- Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047* (2024).
- Haibo Chen, Lei Zhao, Jun Li, and Jian Yang. 2023b. TSSAT: Two-Stage Statistics-Aware Transformation for Artistic Style Transfer. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6878–6887.
- Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1210–1217.
- Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingji Pan, Lei Wang, and Changsheng Xu. 2022. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11326–11336.
- Ming Ding, Wendi Zheng, Wenqi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems* 35 (2022), 16890–16902.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint:2208.01618* (2022).
- Wei Gao, Yijun Li, Yihang Yin, and Ming-Hsuan Yang. 2020. Fast video multi-style transfer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 3222–3230.
- Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

- Gen-2. 2023. Gen-2. Accessed Nov. 1, 2023 [Online] <https://research.runwayml.com/gen2>, <https://research.runwayml.com/gen2>.
- Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2024. Tokenflow: Consistent diffusion features for consistent video editing. In *In International Conference on Learning Representations (ICLR)*.
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2024. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations (ICLR)* (2024).
- Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jimbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940* (2023).
- Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint:2211.13221* (2022).
- Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2023. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133* (2023).
- Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. 2001. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 327–340.
- Jonathan Ho, William Chan, Chittwan Saharia, Jay Whang, Ruiqi Gao, Alexey Grishchenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint:2210.02303* (2022).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 6840–6851.
- Jonathan Ho, Tim Salimans, Alexey Grishchenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022b. Video diffusion models. *arXiv:2204.03458* (2022).
- Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Covideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint:2205.15868* (2022).
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.
- Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023a. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint:2302.09778* (2023).
- Nisha Huang, Yuxin Zhang, and Weiming Dong. 2023b. Style-A-Video: Agile Diffusion for Arbitrary Text-based Video Style Transfer. *arXiv preprint:2305.05464* (2023).
- Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.
- Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, and Yang Gao. 2021. Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14861–14869.
- Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2019. Styling video by example. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–11.
- Levon Khachatryan, Andranik Mojsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shani Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint:2303.13439* (2023).
- Nupur Kumar, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*.
- Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems* 30 (2017).
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.
- Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6649–6658.
- Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2023. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440* (2023).
- Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. 2023. Text-guided synthesis of eulerian cinemagraphs. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–13.
- Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*. PMLR, 8162–8171.
- OpenAI. 2023. GPT-4V(ision) System Card. *Technical report* (2023).
- Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5880–5888.
- Fred Phillips and Brandy Mackintosh. 2011. Wiki Art Gallery, Inc.: A case for critical thinking. *Issues in Accounting Education* 26, 3 (2011), 593–608.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023a. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023b. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint:2307.01952* (2023).
- Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
- Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCP 2016, Hannover, Germany, September 12–15, 2016, Proceedings* 38. Springer, 26–36.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.
- Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411* (2023).
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint:2209.14792* (2022).
- Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. StyleDrop: Text-to-Image Generation in Any Style. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising Diffusion Implicit Models. In *In International Conference on Learning Representations (ICLR)*.
- Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *In International Conference on Learning Representations (ICLR)*.
- Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II* 16. Springer, 402–419.
- Joshua B Tenenbaum and William T Freeman. 2000. Separating style and content with bilinear models. *Neural computation* 12, 6 (2000), 1247–1283.
- Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclie Choi, Sergey Tulyakov, and Daniel Sýkora. 2020a. Arbitrary style transfer using neurally-guided patch-based synthesis. *Computers & Graphics* 87 (2020), 62–71.
- Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclie Choi, Sergey Tulyakov, and Daniel Sýkora. 2020b. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 73–1.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- Bin Wang, Wenping Wang, Huaping Yang, and Jiaguang Sun. 2004. Efficient example-based painting and synthesis of 2d directional texture. *IEEE Transactions on Visualization and Computer Graphics* 10, 3 (2004), 266–277.

- Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023c. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).
- Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2024. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems* 36 (2024).
- Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiahuo Yu, Peiqing Yang, et al. 2023a. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103* (2023).
- Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. 2023b. StyleAdapter: A Single-Pass LoRA-Free Model for Stylized Image Generation. *arXiv preprint:2309.01770* (2023).
- Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15943–15953.
- Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.
- Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2024. ToonCrafter: Generative Cartoon Interpolation. *arXiv preprint arXiv:2405.17933* (2024).
- Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. 2023. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190* (2023).
- Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia 2023 Conference Proceedings*.
- Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2024. FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8703–8712.
- Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2308.06721* (2023).
- Wei Zhang, Chen Cao, Shifeng Chen, Jianzhuang Liu, and Xiaolu Tang. 2013. Style transfer via image component analysis. *IEEE Transactions on multimedia* 15, 7 (2013), 1594–1601.
- Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10156.
- Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.
- Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint:2211.11018* (2022).

Our Supplementary Material consists of 7 sections:

- Section A provides a detailed statement of our experiments, including the implementation details of comparison methods, and details of our test set.
- Section B provides a detailed statement of our evaluation, including the details of evaluation metrics, and details of the user study.
- Section C adds more comparison experiments, including the comparison with StyleDrop, comparison in multi-reference stylized image generation, and comparison with style transfer methods.
- Section D adds additional ablation study on different adapter architecture.
- Section E explores the extended application of StyleCrafter, including the collaboration with depth control.
- Section F demonstrates more results of our methods.
- Section G discusses the limitations.

## A Implementation Details

### A.1 Comparison methods

For all comparison methods, we follow the instructions from the official papers and open-source implementations. Since some methods including Dreambooth and InST require additional finetuning, we provide all implementation details as follows:

*Dreambooth.* Dreambooth [Ruiz et al. 2023] aims to generate images of a specific concept (e.g., style) by finetuning the entire text-to-image model on one or several images. We train Dreambooth based on Stable Diffusion 1.5. The training prompts are obtained from BLIP-2 [Li et al. 2023a], and we manually add a style postfix using the rare token "sks". For example, "two slices of watermelon on a red surface in sks style" is used for the first style reference in Table S3. We train the model for 500 steps for single-reference styles and 1500 steps for multi-reference styles, with learning rates of  $5 \times 10^{-6}$  and a batch size of 1. The training steps are carefully selected to achieve the balance between text alignment and style conformity.

*InST.* InST [Zhang et al. 2023] propose a inversion-based method to achieve style-guided text-to-image generation through learning a textual description from style reference. We train InST for 1000 steps with learning rates of  $1 \times 10^{-4}$  and a batch size of 1.

*StableDiffusion 2.1 and SDXL.* We extend Stable Diffusion to style-guided text-to-video generation by utilizing GPT-4v to generate style descriptions from style reference. Details about style descriptions can be found in Table S3

*IP-Adapter.* IP-Adapter [Ye et al. 2023] propose to train an image-conditioned adapter to generate images from image prompts. We use the official checkpoint of IP-Adapter-Plus(SDXL) for evaluation. Note that IP-Adapter is primarily designed for image variants and other editing tasks. When conducted with its default scale value  $s = 1$ , IP-Adapter tends to simply reconstruct style references, which actually underestimates the ability of IP Adapters in stylized generation. During the evaluation, **we adjust the scale value to 0.5 to ensure a more balanced comparison.**

*Style Aligned.* Style Aligned [Hertz et al. 2023] design a self-attention sharing mechanism to ensure constant style among different samples, supporting both stylized generation and style transfer tasks. We conduct the official implementation on SDXL during the evaluation.

*VideoCrafter and Gen-2.* Similar to SD\*, We use VideoCrafter [Chen et al. 2023a]  $320 \times 512$  Text2Video Model and Gen-2 [Gen-2 2023] equipped with GPT-4v to generate stylized videos from style references and text prompts.

*AnimateDiff.* AnimateDiff [Guo et al. 2024] aims to extend personalized T2I model(i.e., Dreambooth or LoRA [Hu et al. 2022]) for video generation. To compare with AnimateDiff, we first train personalized dreambooth models for each group of multi-reference style images, then we incorporate them into AnimateDiff based on their released codebase. We did not use LoRA because we observed that AnimateDiff fails to turn LoRA-SD for video generation in most cases.

### A.2 Testing Datasets

We provide a detailed description of the testing datasets.

*Content Prompts.* We utilize GPT4 to generate prompts across four meta-categories: human, animal, object, and landscape. Initially, 15/10 prompts (for images/videos) were generated in each category. Recognized as low-quality prompts, semantically repeated prompts, containing style descriptions, and other less informative ones were manually filtered, leading to 5/3 prompts per category. For video prompts, we specifically encouraged the generation of scenarios involving motion. The final prompts in testset are provided in Table S1 and Table S2.

*Style References.* We collect 20 diverse single-reference stylized images and 8 sets of style images with multi-reference(each contains 5 to 7 images in similar styles) from the Internet<sup>1</sup>. Besides, for the comparison with the Text-to-Image model including Stable Diffusion and the Text-to-Video model including VideoCrafter and Gen-2, we extend them to stylized generation by equipped them with GPT-4v to generate textual style descriptions from style reference. We provide style references and corresponding style descriptions in Table S3 and Figure S1.

## B Evaluation Details

### B.1 Evaluation Metrics

We employ CLIP-based similarity scores to evaluate text alignment and style conformity, as is commonly done by existing methods. Additionally, we include two metrics to measure the temporal consistency of generated videos, i.e., *CLIP-Temp* and *Warping Error*. A detailed calculation process for each metric is presented below.

*CLIP-Text.* We utilize the pretrained CLIP-ViT-H-14-laion2B-s32B-b79K<sup>2</sup> as a feature extractor(also for *CLIP-Style* and *CLIP-Temp*), which is trained on LAION-2B and demonstrates enhanced performance across various datasets. We extract the frame-wise image

<sup>1</sup>The style references are collected from <https://unsplash.com/>, <https://unsplash.com/>, <https://en.m.wikipedia.org/wiki/>, <https://civitai.com/>, <https://clipdrop.co/>

<sup>2</sup><https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K>

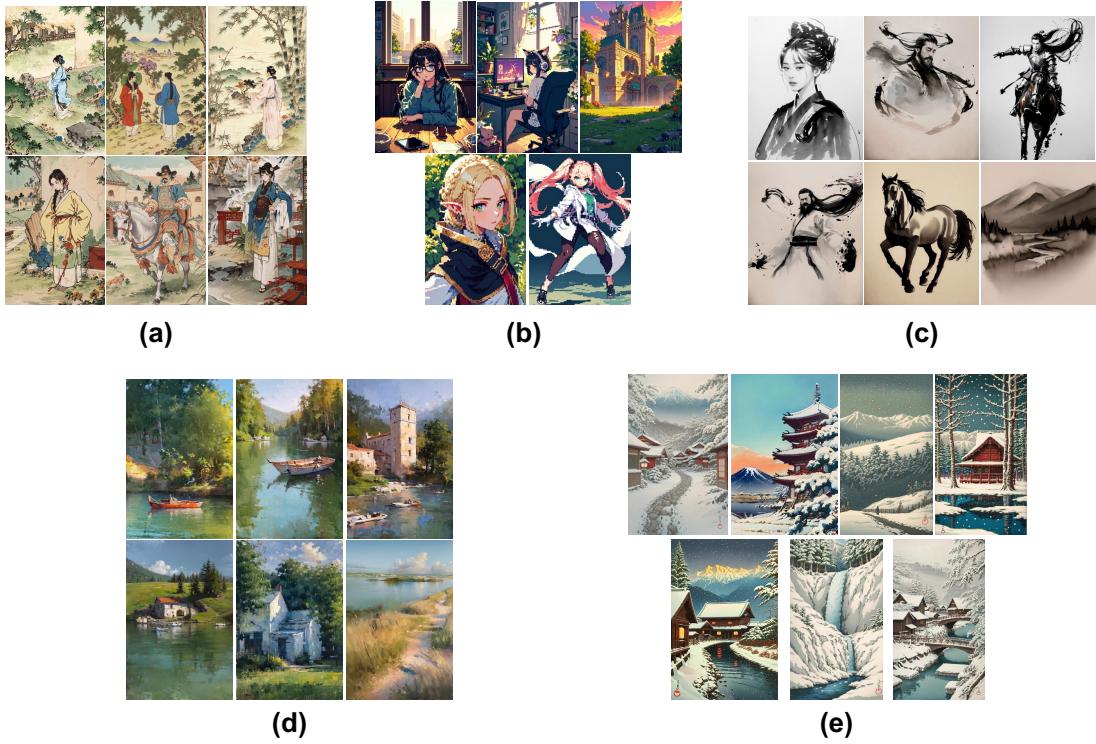


Fig. S1. Multiple references in the testset

Table S1. Text prompts used in the testset for image generation

Prompt	Meta Category	Prompt	Meta Category
A man playing the guitar on a city street.	Human	A flock of birds flying gracefully in the sky.	Animal
A woman reading a book in a park.	Human	A colorful butterfly resting on a flower.	Animal
A couple dancing gracefully together.	Human	A bear fishing in a river.	Animal
A person sitting on a bench, feeding birds.	Human	A dog running in front of a house.	Animal
A person jogging along a scenic trail.	Human	A rabbit nibbling on a carrot.	Animal
A bouquet of flowers in a vase.	Object	A cobblestone street lined with shops and cafes.	Landscape
A telescope pointed at the stars.	Object	A modern cityscape with towering skyscrapers.	Landscape
A rowboat docked on a peaceful lake.	Object	A winding path through a tranquil garden.	Landscape
A lighthouse standing tall on a rocky coast.	Object	An ancient temple surrounded by lush vegetation.	Landscape
A rustic windmill in a field.	Object	A serene mountain landscape with a river flowing through it.	Landscape

Table S2. Text prompts used in the testset for video generation

Prompt	Meta Category	Prompt	Meta Category
A street performer playing the guitar.	Human	A bear catching fish in a river.	Animal
A chef preparing meals in kitchen.	Human	A knight riding a horse through a field.	Animal
A student walking to school with backpack.	Human	A wolf walking stealthily through the forest.	Animal
A campfire surrounded by tents.	Object	A river flowing gently under a bridge.	Landscape
A hot air balloon floating in the sky.	Object	A field of sunflowers on a sunny day.	Landscape
A rocketship heading towards the moon.	Object	A wooden sailboat docked in a harbor.	Landscape

## Stylized Video Generation -- User Study

Stylized Video Generation is an extension of the Text-to-Video Generation, designed to generate content-controlled and style-controlled videos from a given style reference and content description.

Watch the following video results generated from the style reference and text description, with 3 sub-questions for each set of comparisons (please **separately** review the generated results from the following three perspectives:

1. **Text Alignment** (**multiple choice**, select **all** results that is aligned with the text description);
2. **Style Conformity** (**single choice**, select the result that is closest to the reference image **in style**);
3. **Temporal Quality** (**single choice**, select the result with the best temporal quality);

Notes:

1. **Text Alignment** means that the **content** of the generated video is aligned with the **text description** (prompt), and the content of the text description should appear in the generated result;
  2. **Style Conformity** means that the **style** of the generated video is consistent with the style of the **reference image**, where the style includes both the color tone, texture, brush strokes, etc., as well as the painting style, emotion, and mood;
  3. **Temporal Quality** consists of two aspects: First, the generated video should **include** certain **action** or **camera movement**, and should be in line with the picture context; Second, the content of the picture should be **coherent**, **without abrupt changes** or flickering;
4. Please ignore the watermark effect and the missing area in the bottom right corner of the result.



\* 1.1 Which one is aligned with **text description?** [Multiple choice]

- A       B       C

\* 1.2 Which one performs best in **Style Conformity?**

- A       B       C

\* 1.3 Which one performs best in **Temporal Quality?**

- A       B       C

Fig. S2. User Preference Study Interface

Table S3. Style references in the testset and corresponding style descriptions generated from GPT-4v[OpenAI 2023].

Style Reference	Style Descriptions	Style Reference	Style Descriptions
	3D Digital Art, <a href="#">[prompt]</a> , whimsical and modern, smooth and polished surfaces, bold and contrasting colors, soft shading and lighting, surreal representation.		Digital Painting, <a href="#">[prompt]</a> , detailed rendering, vibrant color palette, smooth gradients, realistic light and reflection, immersive natural landscape scene.
	Manga-inspired digital art, <a href="#">[prompt]</a> , dynamic composition, exaggerated proportions, sharp lines, cel-shading, high-contrast colors with a focus on sepia tones and blues.		Childlike watercolor, <a href="#">[prompt]</a> , simple brush strokes, primary and secondary colors, bold outlines, flat washes, playful, spontaneous, and expressive.
	Comic book illustration, <a href="#">[prompt]</a> , digital medium, clean inking, cell shading, saturated colors with a natural palette, and a detailed, textured background.		Pixel art illustration, <a href="#">[prompt]</a> , digital medium, detailed sprite work, vibrant color palette, smooth shading, and a nostalgic, retro video game aesthetic.
	Ink and watercolor on paper, <a href="#">[prompt]</a> , urban sketching style, detailed line work, washed colors, realistic shading, and a vintage feel.		Flat Vector Illustration, <a href="#">[prompt]</a> , simplified shapes, uniform color fills, minimal shading, absence of texture, clean and modern aesthetic.
	Watercolor and ink illustration, <a href="#">[prompt]</a> , traditional comic style, muted earthy color palette, detailed with a sense of movement, soft shading, and a historic ambiance.		Low Poly Digital Art, <a href="#">[prompt]</a> , geometric shapes, vibrant colors, flat texture, sharp edges, gradient shading, modern graphic style.
	Chinese ink wash painting, <a href="#">[prompt]</a> , minimalist color use, calligraphic brushwork, emphasis on flow and balance, with poetic inscription.		Chinese Ink Wash Painting, <a href="#">[prompt]</a> , monochromatic palette, dynamic brushstrokes, calligraphic lines, with a focus on negative space and movement.
	Manga Style, <a href="#">[prompt]</a> , black and white digital inking, high contrast, detailed line work, cross-hatching for shadows, clean, no color.		Line Drawing, <a href="#">[prompt]</a> , simple and clean lines, monochrome palette, smooth texture, minimalist and cartoonish representation .
	Van Gogh's "Starry Night" style, <a href="#">[prompt]</a> , with expressive, swirling brushstrokes, rich blue and yellow palette, and bold, impasto texture.		Watercolor Painting, <a href="#">[prompt]</a> , fluid brushstrokes, transparent washes, color blending, visible paper texture, impressionistic style.
	Van Gogh-inspired pen sketch, <a href="#">[prompt]</a> , dynamic and swirling line work, monochromatic sepia tones, textured with a sense of movement and energy.		Ukiyo-e Woodblock Print, <a href="#">[prompt]</a> , gradation, limited color palette, flat areas of color, expressive line work, stylized wave forms, traditional Japanese art.
	Watercolor Painting, <a href="#">[prompt]</a> , fluid washes of color, wet-on-wet technique, vibrant hues, soft texture, impressionistic portrayal.		Victorian watercolor, <a href="#">[prompt]</a> , fine detail, soft pastel hues, gentle lighting, clear texture, with a quaint, realistic portrayal of everyday life.

embeddings from the generated results and text embeddings from the input content prompts, then compute their average cosine similarity. The overall *CLIP-Text* is calculated as:

$$S_{text} = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{T} \sum_{t=1}^T \frac{\text{emb}(x_t^i) \cdot \text{emb}(p^i)}{\|\text{emb}(x_t^i)\| \cdot \|\text{emb}(p^i)\|} \right) \quad (3)$$

where  $M$  represents the total number of testing videos and  $T$  represents the total number of frames in each video ( $T = 1$  for image generation),  $\text{emb}(x_t^i)$  and  $\text{emb}(p^i)$  indicate the CLIP embedding of the  $t$ -th frame of the  $i$ -th video  $x_t^i$  and the corresponding prompt  $p^i$ , respectively.

*CLIP-Style*. Similarly, we extract the frame-wise image embeddings  $\text{emb}(x_t^i)$  from the generated results and image embeddings  $s^i$  from the input style reference. The overall *CLIP-Text* is calculated as:

$$S_{style} = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{T} \sum_{t=1}^T \frac{\text{emb}(x_t^i) \cdot \text{emb}(s^i)}{\|\text{emb}(x_t^i)\| \cdot \|\text{emb}(s^i)\|} \right) \quad (4)$$

*CLIP-Temp*. Considering the semantic consistency between every two frames, we extract the frame-wise CLIP image embeddings and compute the cosine similarity between each of the two frames, as follows:

$$S_{temp} = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\text{emb}(x_t^i) \cdot \text{emb}(x_{t+1}^i)}{\|\text{emb}(x_t^i)\| \cdot \|\text{emb}(x_{t+1}^i)\|} \right) \quad (5)$$

*Warping Error*. For the warping error, we first obtain the optical flow between each two frames using RAFT-small [Teed and Deng 2020], a pre-trained optical flow estimation network. Subsequently, we compute the pixel-wise differences between the warped image and the predicted image, as follows:

$$W_{error} = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{T-1} \sum_{t=1}^{T-1} \|x_t^i - \text{warp}(x_{t+1}^i, \text{flow}(x_{t+1}^i, x_t^i))\| \right) \quad (6)$$

where  $\text{warp}(x_{t+1}^i, \text{flow}(x_{t+1}^i, x_t^i))$  represents the warped frame of  $x_{t+1}^i$  using the optical flow between frame  $x_{t+1}^i$  and frame  $x_t^i$ .

## B.2 User Study

In this subsection, we provide a detailed introduction about our user study. We randomly selected 15 single-reference style-text pairs to compare the generated results among VideoCrafter [Chen et al. 2023a], Gen-2 [Gen-2 2023], and our proposed method. Given that videocomposer [Wang et al. 2024] directly replicates the style reference and is minimally influenced by the prompt in most cases, we excluded it from the comparison in the user study. Additionally, we randomly chose 10 multi-reference style-text pairs for the comparison between AnimateDiff [Guo et al. 2024] (multiple style-specific models) and our method (a generic model). To ensure a blind comparison, we randomized the order of options for each question and masked the possible model watermark in the lower right corner.

The designed user preference interface is illustrated in Figure S2. We invited 15 users of normal eyesight to evaluate the generated

results in three aspects: text alignment, style conformity, and temporal quality. The instructions and questions are provided as below. Consequently, a total of 1125 votes are collected.

### Instructions.

- **Task:** Watch the following video results generated from the style reference and text description, with 3 sub-questions for each set of comparisons (please **separately** review the generated results from the following three perspectives):
  - **Text Alignment** (multiple choice, means that the content of the generated video is aligned with the text description(prompt), and the content of the text description should appear in the generated result);
  - **Style Conformity** (single choice, means that the style of the generated video is consistent with the style of the reference image, where the style includes both the color tone, texture, brush strokes, etc., as well as the painting style, emotion, and mood);
  - **Temporal Quality** (single choice, consists of two aspects: First, the generated video should include certain action or camera movement, and should be in line with the picture context; Second, the content of the picture should be coherent, without abrupt changes or flickering);
- Please ignore the watermark effect and the missing area in the bottom right corner of the result.

### Questions.

- Which one is aligned with text description? [**Multiple choice**]
- Which performs best in Style Conformity? [**Single choice**]
- Which performs best in Temporal Quality? [**Single choice**]

## C Extended Comparison

### C.1 Multi-reference Stylized Image Generation

We conduct comparisons of multi-reference stylized image generation with Dreambooth [Ruiz et al. 2023] and CustomDiffusion [Kumari et al. 2023], both of which support generating images in specific styles by finetuning on the reference images. Figure S1 and Table S4 present the visual and quantitative results respectively, demonstrating that our method surpasses all competitors in terms of style conformity for multi-reference stylized generation. Although Dreambooth and CustomDiffusion exhibit competitive performance in certain cases, their stylized generation abilities tend to vary with different prompts, i.e., struggling to maintain consistent visual styles across arbitrary prompts. It is possibly because several images are insufficient to allow the model to disentangle the contents and styles, thus harming the generalization performance. Besides, the requirement for finetuning during the testing process also undermines their flexibility. In contrast, our method efficiently generates high-quality stylized images that align with the prompts and conform the style of reference images without additional finetuning costs.

Table S4. Quantitative comparison on Multi-reference style-guided T2I generation. **Bold**: Best.

Methods	Dreambooth	CustomDiffusion	Ours
Text ↑	0.2868	<b>0.2986</b>	0.2924
Style ↑	0.4270	0.4441	<b>0.5333</b>

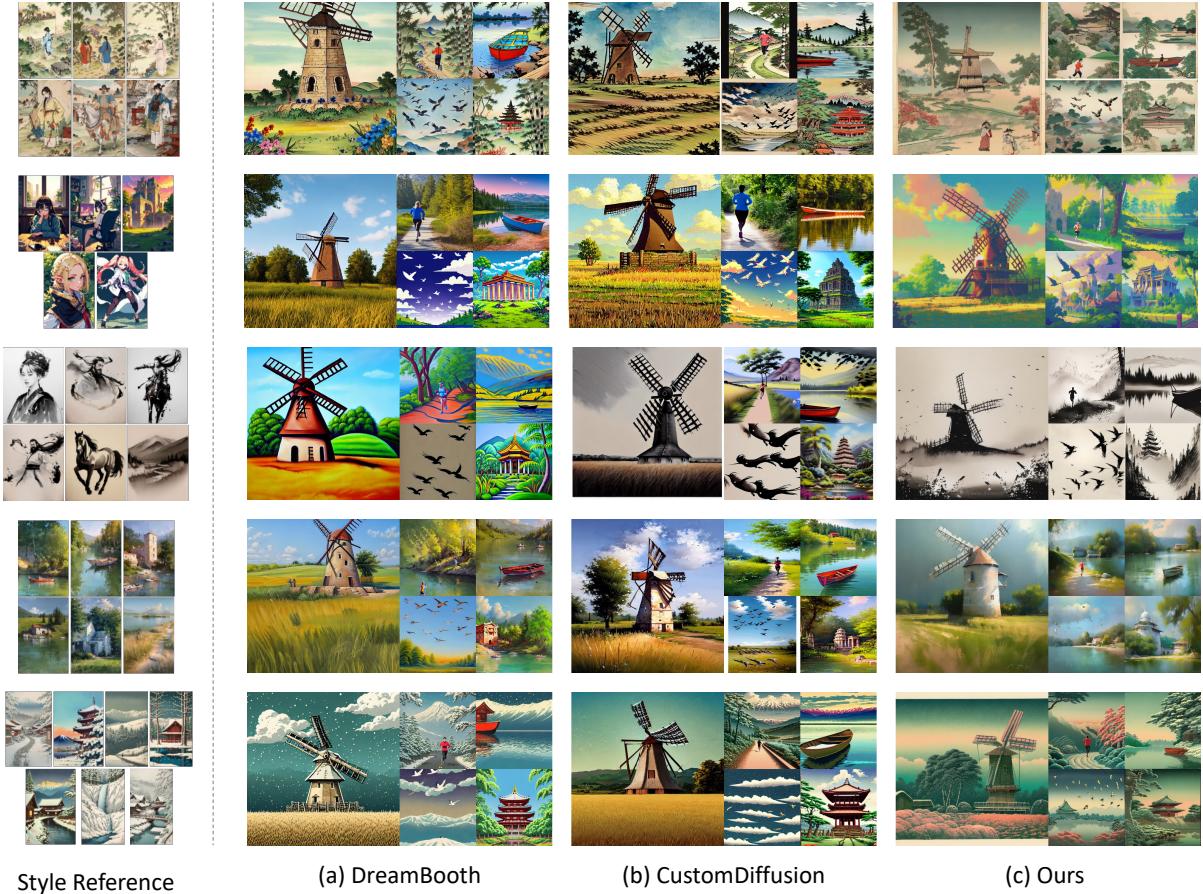


Fig. S3. Visual comparison on multi-reference stylized T2I generation. Testing prompts: (i) A rustic windmill in a field; (ii) A person jogging along a scenic trail; (iii) A flock of birds flying gracefully in the sky; (iv) A rowboat docked on a peaceful lake; (v) An ancient temple surrounded by lush vegetation.

## C.2 Comparison with StyleDrop

Here we present a supplementary comparison with StyleDrop [Sohn et al. 2023]. StyleDrop proposes a versatile method for synthesizing images that faithfully follow a specific style using a text-to-image model. Owing to the absence of an official StyleDrop implementation, we have excluded the comparison with StyleDrop from the main text. Instead, we include a comparison with an unofficial StyleDrop implementation<sup>3</sup> here. We train StyleDrop based on StableDiffusion 2.1 for 1000 steps with a batch size of 8 and a learning rate of  $3 \times 10^{-4}$ . The quantitative and qualitative results are presented in Table S5 and Figure S5 respectively. Results show that compared to StyleDrop, our proposed method more effectively captures the visual characteristics of a user-provided style and combines them with various prompts in a flexible manner.

Table S5. Quantitative comparison with StyleDrop.

Methods	StyleDrop	Ours(SD 2.1)
Text ↑	0.2389	0.3028
Style ↑	0.3962	0.4836

<sup>3</sup><https://github.com/aim-uofa/StyleDrop-PyTorch>

## C.3 Comparison with Style Transfer Methods

In this section, we perform comparisons with a particular set of competitors, i.e., style transfer methods. As style transfer methods require extra source images or videos for content provision, we utilize text-to-image models(Stable Diffusion 2.1) and text-to-video models(VideoCrafter) to initially generate visual content. Subsequently, we apply existing style transfer methods based on the reference to produce the stylized output. For stylized image generation, we opt for T2I + CAST[Zhang et al. 2022], while for stylized video generation, we choose T2V + MCCNet[Deng et al. 2021]. The results of this comparison are illustrated in the Table S6 and Table S7. Both two competitors underperform our approach. The reasons are in two aspects: (i) Style transfer methods mainly work well on transferring tones and local textures while fall short in transferring semantic style features; (ii) Style transfer method cannot change the structure of the content image, which hinders the generation of styles that associated with special geometry features, e.g. layouts of logo style and 3D render style.

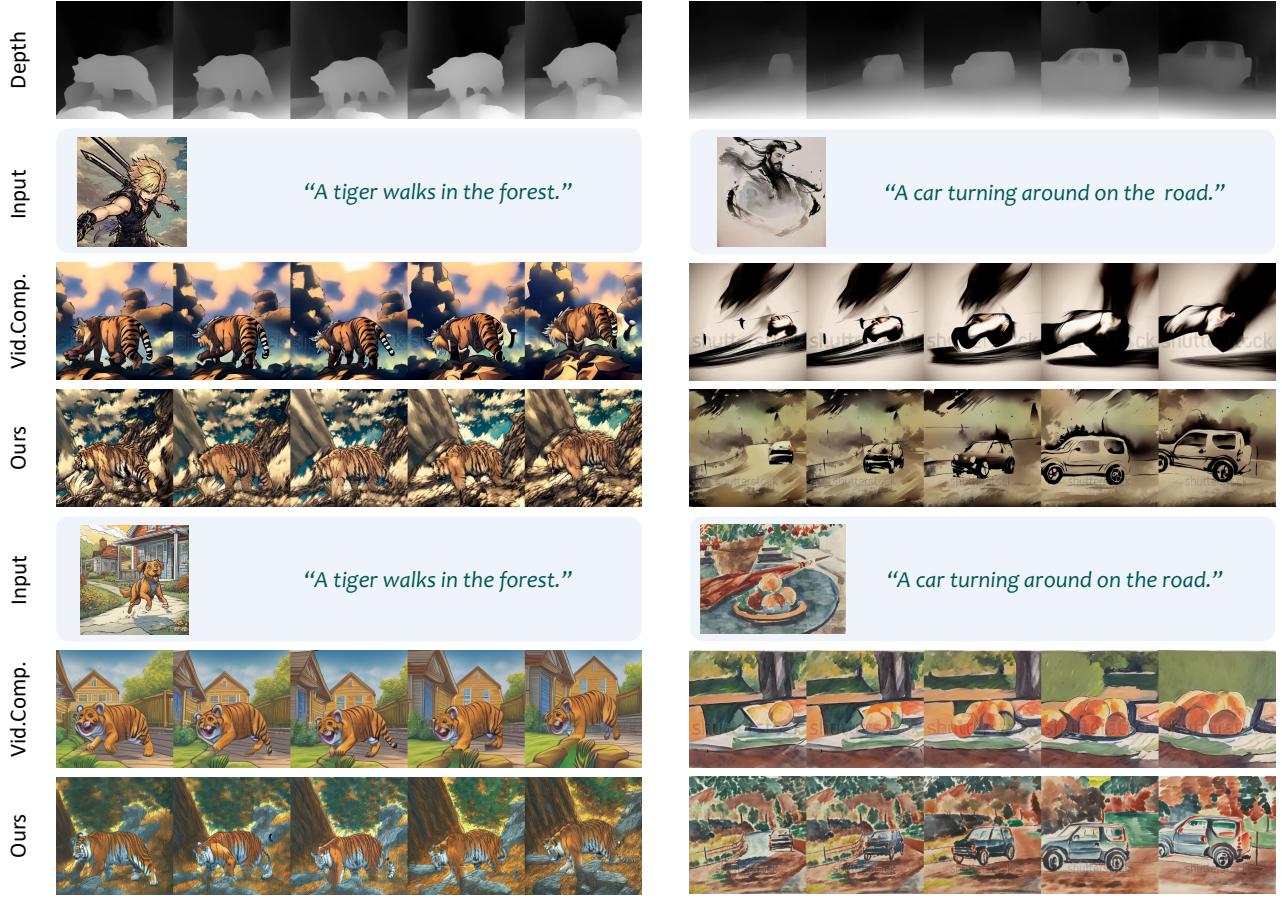


Fig. S4. Visual comparison on stylized video generation with additional depth guidance. Vid.Comp.: VideoComposer



Fig. S5. Visual comparison between StyleDrop and our proposed method. Testing prompts: (i) A woman reading a book in a park.; (ii) A person jogging along a scenic trail.; (iii) A colorful butterfly resting on a flower.; (iv) A rabbit nibbling on a carrot.; (v) A telescope pointed at the stars.

## D Extended Ablation Study

Since we have made ablation studies on some key design in the main text(i.e., Dual Cross-Attention, Data Augmentation, and Adaptive

Table S6. Quantitative comparison with image style transfer method.

Methods	T2I + CAST	Ours(SD 2.1)
Text ↑	0.3027	0.3028
CLIP-Style ↑	0.3549	0.4836

Table S7. Quantitative comparison with video style transfer method.

Methods	T2I + MCCNet	Ours(VideoCrafter)
Text ↑	0.2487	0.2726
Style ↑	0.2858	0.4531
Temporal ↑	0.9577	0.9892

Style-Content Fusion), we provide additional comparison between different style adapter architectures, including MLP, Transformer, and Query Transformer(ours). Quantitative results and visual comparisons in Figure S6 and Table S8 show that Query Transformer excels over the other alternatives.

Table S8. Ablation studies on style feature extractor architecture. The performance is evaluated on the style-guided T2I generation.

Alternatives	MLP	Transformer	Q-Former (Ours)
Text ↑	0.3415	0.3221	0.3028
Style ↑	0.2843	0.4149	0.4836

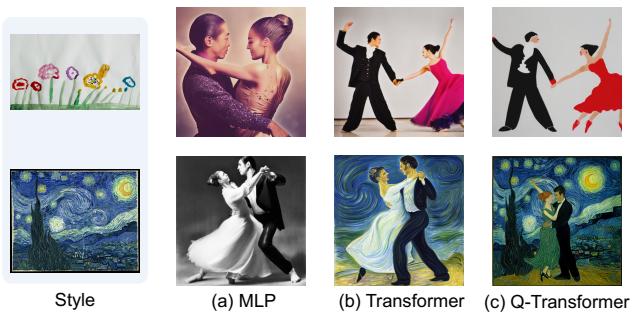


Fig. S6. Visual comparison between different adapter architectures. Prompt: *A couple dancing gracefully together.*

## E Application Extension

In this section, we further explore the compatibility with additional controllable conditions, e.t., depth. Following the approach of structure control in Animate-A-Story[He et al. 2023], we introduce video structure control by integrating a well-trained depth adapter into the base T2V model. Note that StyleCrafter and depth-adaptor are trained independently, the only operation we take is to combine the both during the inference stage. Instead of employing DDIM Inversion to ensure consistency, we generate the videos from random noise. The visual comparison with VideoComposer[Wang et al. 2024] is present in Figure S4. VideoComposer struggles to produce results faithful to text descriptions when faced with artistic styles, such as the "boat" mentioned in the prompt. In contrast, our method not only supports collaboration with depth guidance, but also generates videos with controllable content, style, and structure.

## F More Results

In this section, we provide more visual results and comparisons of our method. Specifically, we provide: (i) style-guided text-to-image generation results on StyleCrafter(SD2) and StyleCrafter(SDXL) in Figure S8 and Figure S9; (ii) comparison of single-reference stylized video generation and multi-reference stylized video generation, as illustrated in Figure S10 and Figure S11, respectively. (iii) additional stylized video results in Figure S12 and Figure S13.

## G Limitations

While our proposed method effectively handles most common styles, it does have certain limitations. Firstly, since StyleCrafter is developed based on existing T2I models and T2V models, such as SDXL and VideoCrafter, it unavoidably inherits part of the base model's shortcomings, such as fixed resolution and video length, less satisfactory temporal consistency. For example, our method fails to

generate high-definition faces in certain cases, as shown in Figure S7. Despite the fact that our approach successfully enhances the stylistic generation capacity of T2I/T2V models, leveraging a powerful base model will always amplify its performance, as showcased in the superior style conformity of StyleCrafter(SDXL) over StyleCrafter(SD2).

Besides, artistic style is a very comprehensive perceptual feeling and visual styles are considerably more complex than what we explore in our paper. Our model may produce just passable results when confronted with reference images possessing highly stylized semantics. For example, as depicted in Figure S7, although our model successfully reproduces ink strokes, there are still discrepancies with reference images in the aesthetic level, such as the lack of "blank-leaving" in the generation results. Additionally, considering the absence of stylized video data, our stylized video generation results are somewhat less satisfactory than stylized image generation in visual style expression. A possible solution is to collect sufficient stylized video data for training, which we leave for further work.

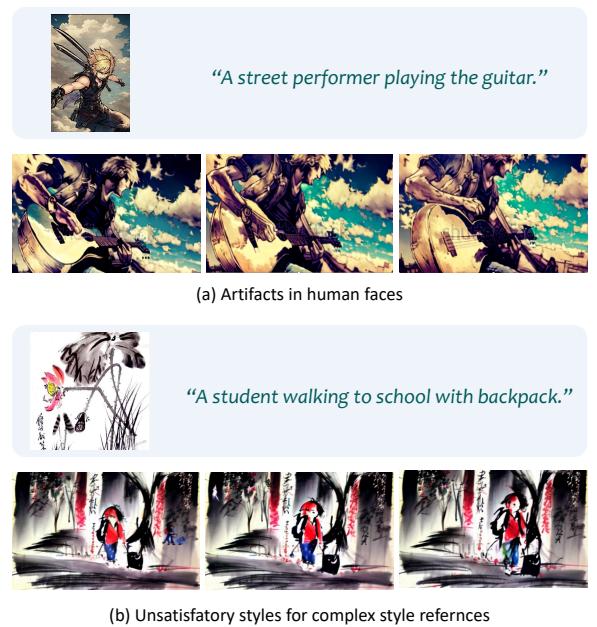


Fig. S7. Failure cases of our methods. (a). Our method inherits limitations from the pretrained T2V model, tends to generate human faces with obvious artifacts. (b). Our method produces less satisfactory styles for complex style references such as Chinese Ink Painting.

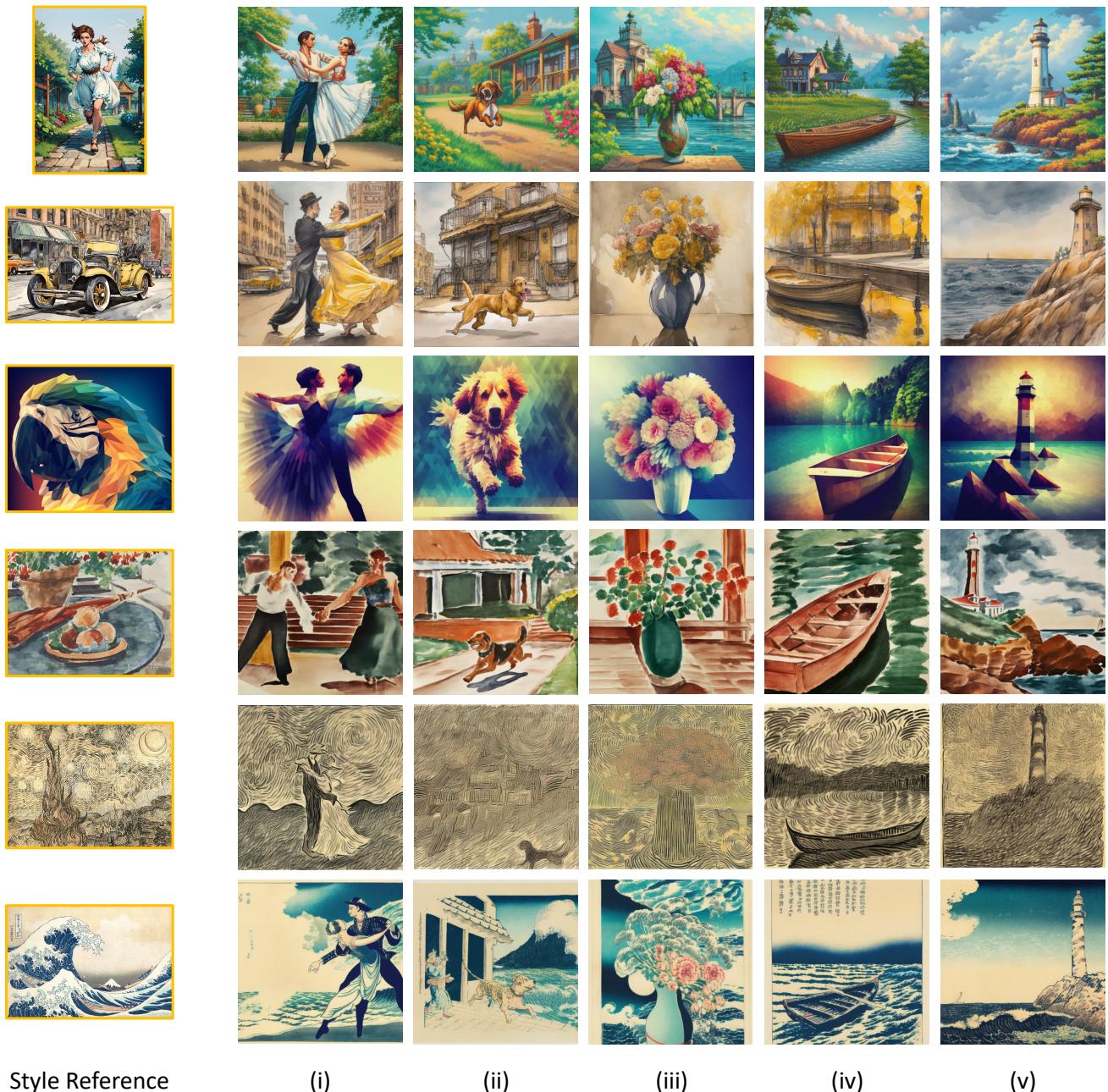
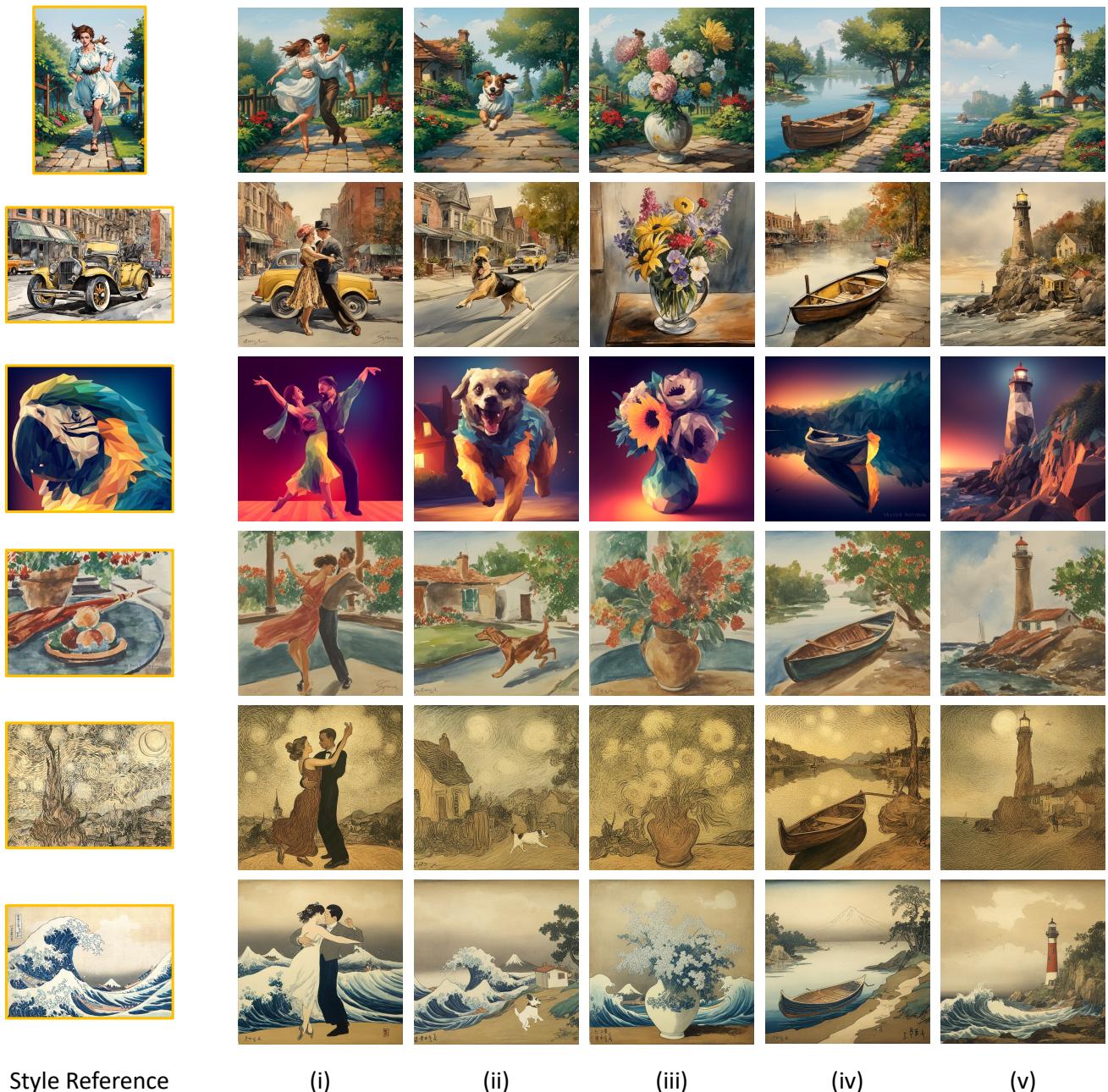


Fig. S8. More Results of **StyleCrafter(SD2)** on Style-Guided Text-to-Image Generation. Prompts: (i) A couple dancing gracefully together. (ii) A dog is running in front of a house. (iii) A bouquet of flowers in a vase. (iv) A rowboat docked on a peaceful lake. (v) A lighthouse standing tall on a rocky coast.



Style Reference

(i)

(ii)

(iii)

(iv)

(v)

Fig. S9. More Results of **StyleCrafter(SDXL)** on Style-Guided Text-to-Image Generation. Prompts: (i) A couple dancing gracefully together. (ii) A dog is running in front of a house. (iii) A bouquet of flowers in a vase. (iv) A rowboat docked on a peaceful lake. (v) A lighthouse standing tall on a rocky coast.

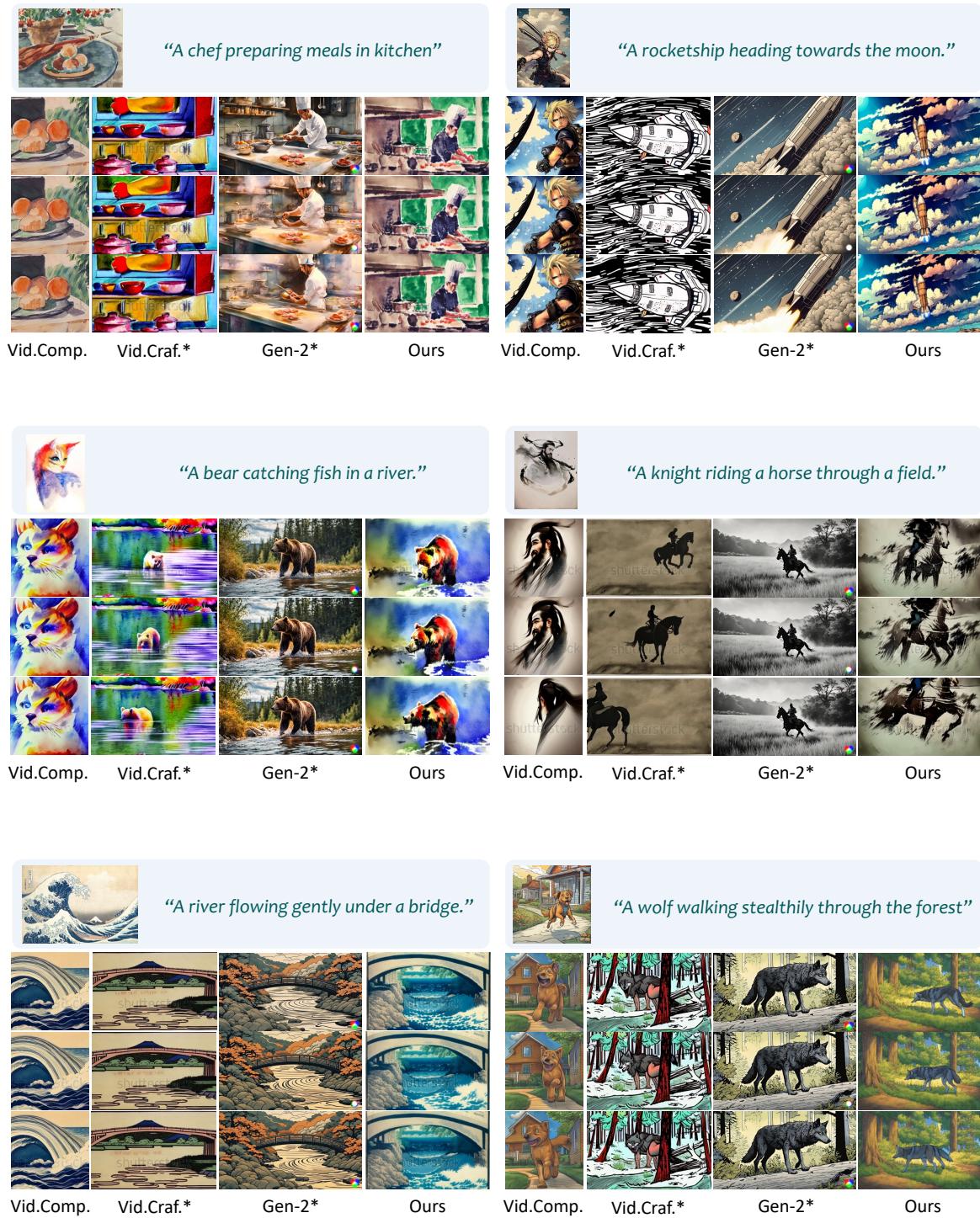


Fig. S10. More Visual Comparison on Single-Reference Stylized T2V Generation. Vid.Comp.: VideoComposer; Vid.Craf.: VideoCrafter.

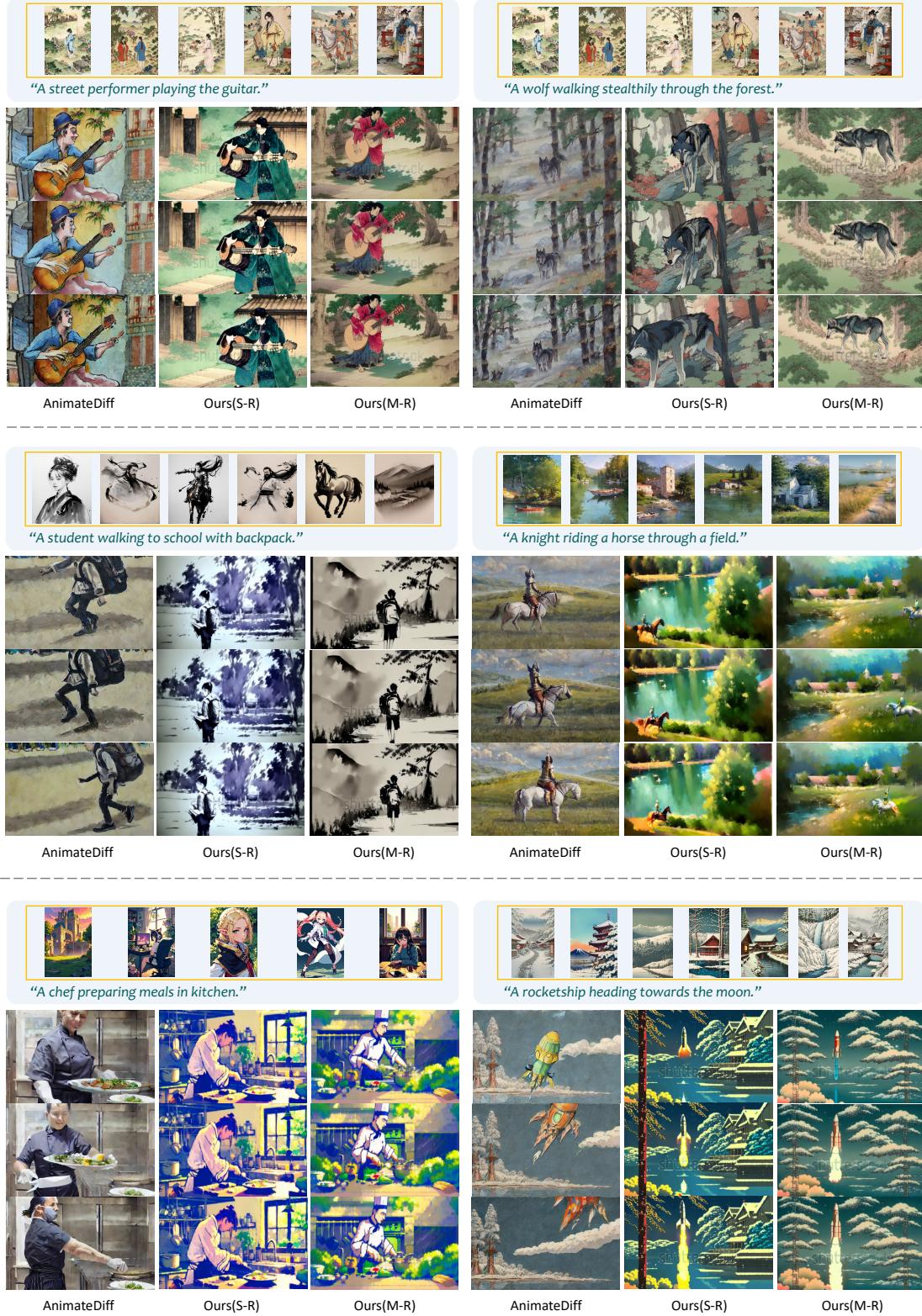


Fig. S11. More Visual Comparison on Multi-Reference Stylized T2V Generation

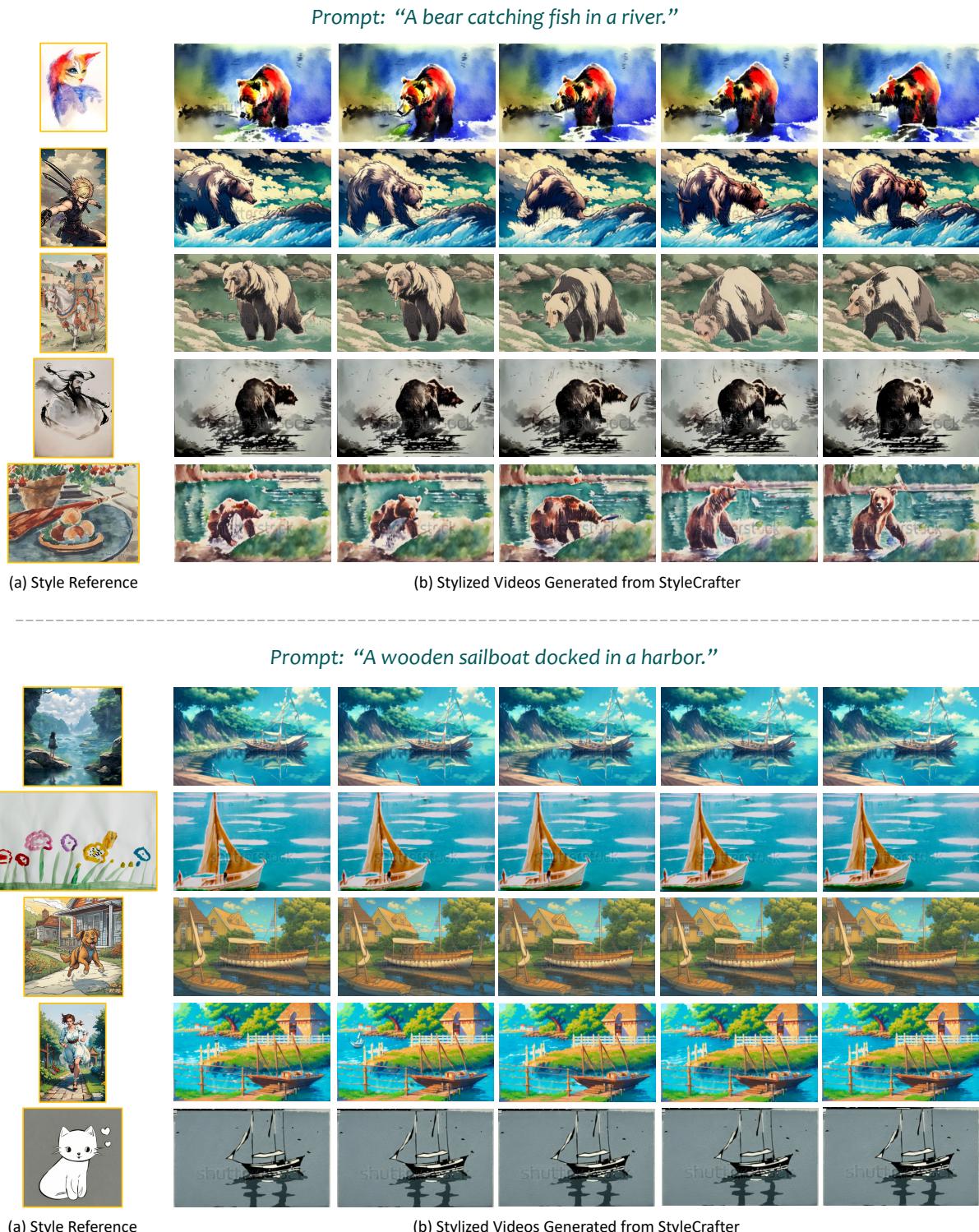
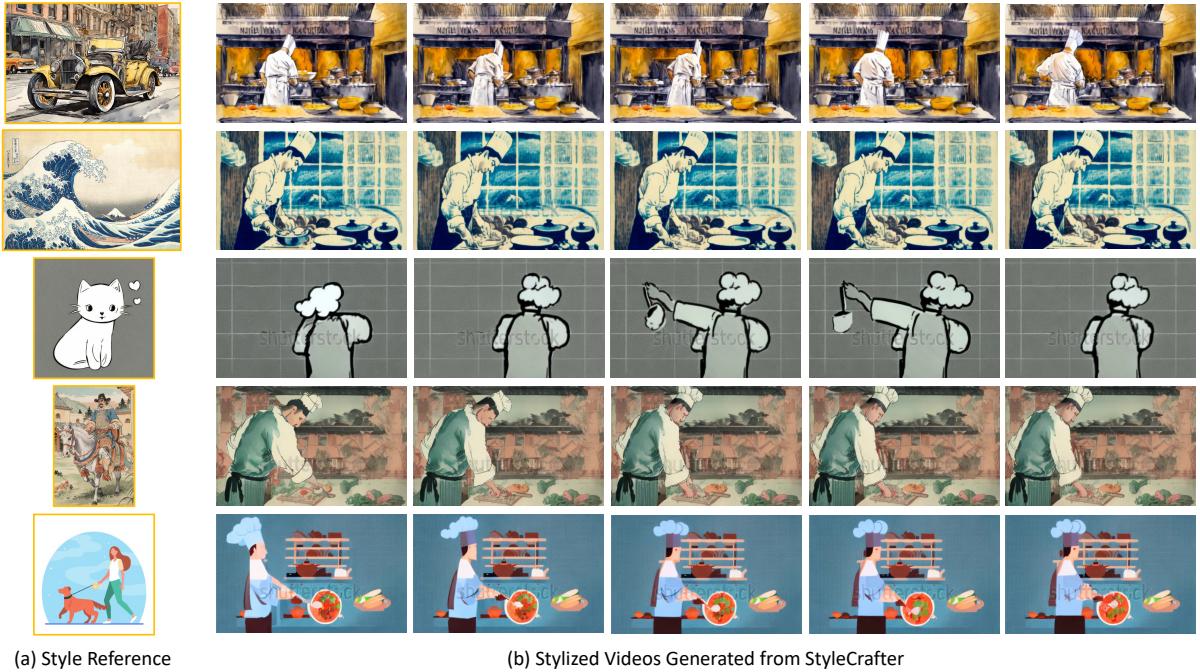


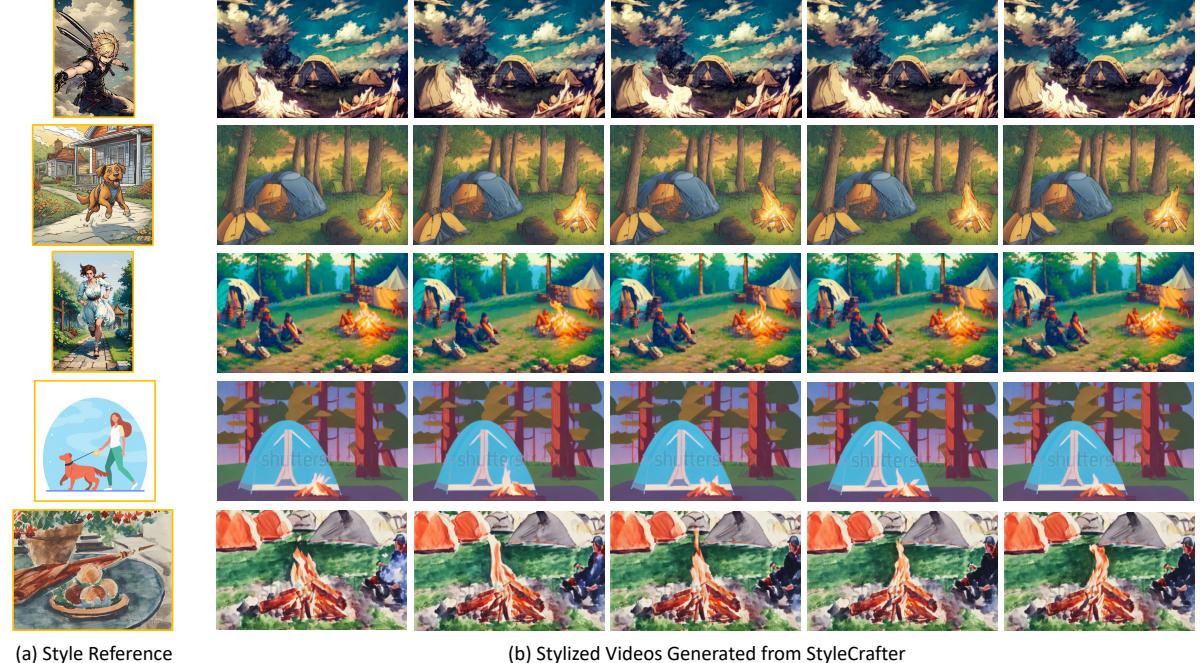
Fig. S12. More Results of StyleCrafter on Style-Guided Text-to-Video Generation

*Prompt: “A chef preparing meals in kitchen.”*



(b) Stylized Videos Generated from StyleCrafter

*Prompt: “A campfire surrounded by tents.”*



**Fig. S13.** More Results of StyleCrafter on Style-Guided Text-to-Video Generation