

DreamStyle: A Unified Framework for Video Stylization

Mengtian Li^{*} Jinshu Chen Songtao Zhao[†] Wanquan Feng Pengqi Tu Qian He
 Intelligent Creation, ByteDance

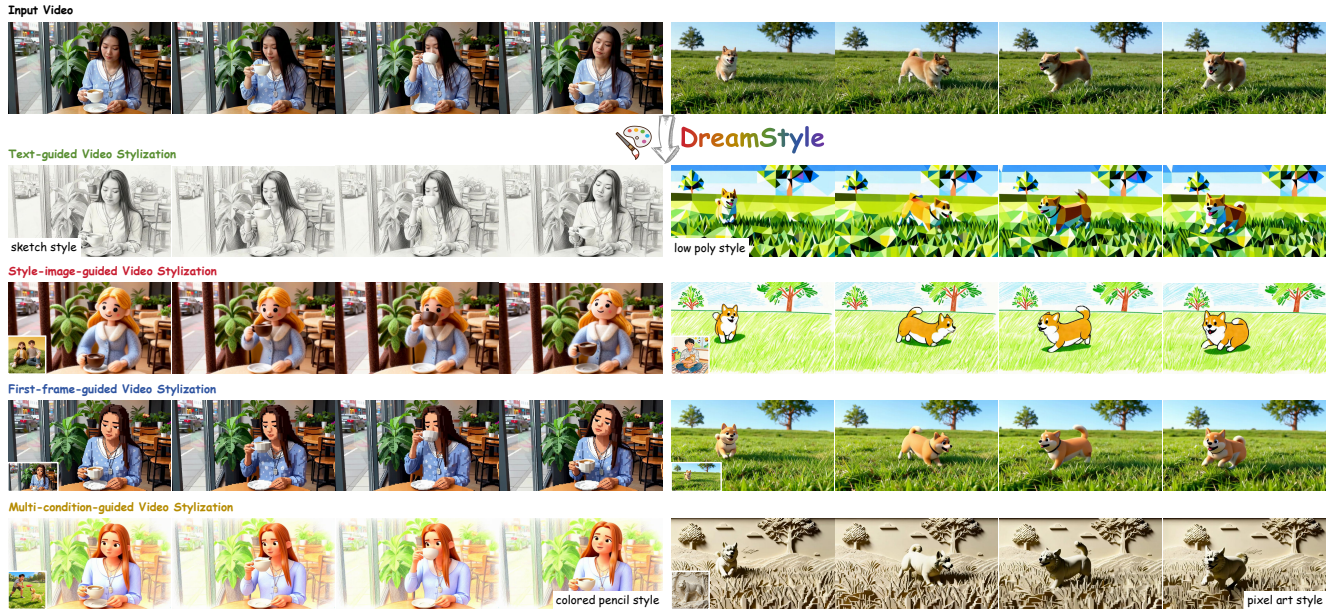


Figure 1. We propose DreamStyle, a unified video stylization framework, which provides a flexible and practical tool for users to create high-quality stylized videos. Given an input video and the reference styles in forms of text, style image, or stylized first frame, DreamStyle faithfully generates videos that align with the desired style—while preserving the main content of the input video.

Abstract

Video stylization, an important downstream task of video generation models, has not yet been thoroughly explored. Its input style conditions typically include text, style image, and stylized first frame. Each condition has a characteristic advantage: text is more flexible, style image provides a more accurate visual anchor, and stylized first frame makes long-video stylization feasible. However, existing methods are largely confined to a single type of style condition, which limits their scope of application. Additionally, their lack of high-quality datasets leads to style inconsistency and temporal flicker. To address these limitations, we introduce **DreamStyle**, a unified framework for video stylization, supporting (1) text-guided, (2) style-image-guided, and (3) first-frame-guided video stylization, accompanied by a well-designed data curation pipeline to acquire high-

quality paired video data. DreamStyle is built on a vanilla Image-to-Video (I2V) model and trained using a Low-Rank Adaptation (LoRA) with token-specific up matrices that reduces the confusion among different condition tokens. Both qualitative and quantitative evaluations demonstrate that DreamStyle is competent in all three video stylization tasks, and outperforms the competitors in style consistency and video quality.

1. Introduction

Video stylization stands as a compelling yet challenging task in the field of visual content generation. Existing video stylization approaches are confronted with the following critical limitations: **(1) Limited stylization capabilities due to single-modality condition.** Text prompts and style images are the two dominant style conditions, but both suffer from inherent flaws. Text prompts are typically ambiguous and unconstrained, failing to precisely describe

^{*}Corresponding author

[†]Project Lead

most abstract styles. Style images, while more visually accurate, exhibit inferior user-friendliness, flexibility, and creativity—it is difficult to acquire a suitable style image, especially for unseen styles. Consequently, most existing methods are confined to styles that are either explicitly describable via text or have clear visual references, exhibiting limited generalization to novel styles. **(2) Scarcity of high-quality modality-aligned training data.** Some existing methods [5, 26, 56] acquire stylization capabilities from image stylization datasets and subsequently generalize to the video domain assisted by a pre-trained video generation model. This paradigm inherently introduces an unavoidable trade-off among style consistency, temporal consistency, and motion dynamics. More recently, UNIC [55] synthesizes stylized videos via a Text-to-Video (T2V) model and employs a gray tiled ControlNet to invert these stylized videos into their realistic counterparts, thereby constructing paired video data. However, its stylization quality is limited by the T2V model and it fails to handle the styles involving geometric deformation due to the strict alignment of tile ControlNet. **(3) Insufficient exploration of potential extended applications.** Current research predominantly focuses on basic stylization capabilities, with limited attention to high-demand extended scenarios, such as multi-style fusion and long-video stylization.

To tackle the aforementioned challenges, we propose **DreamStyle**, which includes the following three key innovations: First, we introduce a unified Video-to-Video (V2V) stylization framework, which is built upon a vanilla I2V model. Through a meticulous design of the condition injection mechanism, we manage to unify diverse forms of style guidance including text prompt, style image, and stylized first frame into a single model, extending the I2V base model to V2V domain while preserving its original architecture and inherent capabilities. We further employ a modified LoRA module composed of a shared down matrix and token-specific up matrices to enhance the multi-task adaptability. Second, we present a systematic data curation pipeline tailored for the video stylization task, the core of which involves two steps: (1) stylizing the initial frame of a real-world video using image stylization techniques and (2) generating the full stylized video sequence from the stylized first frame via an I2V model equipped with ControlNets. To guarantee the data quality, we further adopt a hybrid filtering strategy consisting of automatic and manual filtering. Leveraging this pipeline, we construct two datasets with distinct scales and quality to facilitate multi-stage training. Finally, comprehensive evaluations across multiple dimensions exhibit that our **unified** model, DreamStyle, achieves competitive performance against **specialized** models across various video stylization tasks. Notably, we also demonstrate that allowing multiple style conditions within a single forward process is a crucial design for improving the ef-

fectiveness and controllability of video stylization—such a design unlocks the model’s capability to support more potential extended applications, such as multi-style fusion and long-video stylization.

Overall, our contributions are summarized as follows.

Paradigm. We introduce DreamStyle, which consists of a unified framework that supports text-guided, style-image-guided, and first-frame-guided video stylization; a well-designed pipeline for constructing high-quality paired data for video stylization.

Technology. DreamStyle framework presents a condition injection mechanism that enables seamless handling of diverse video stylization tasks within a unified model; a novel LoRA module that mitigates the interference among different condition tokens.

Scalability. DreamStyle data curation pipeline is practical and scalable for video stylization, overcoming the scarcity of high-quality data and the inherent trade-off between style fidelity and temporal coherence.

Significance. DreamStyle outperforms specialized competitors in various video stylization tasks and exhibits the potential for under-explored extended tasks.

2. Related Work

2.1. Video Diffusion Model

Diffusion models [7, 15, 29, 40, 43] have driven remarkable advancements in visual content generation. Latent Diffusion Models [33, 37] (LDMs) further optimize this paradigm by training a diffusion network in the latent space of pretrained Variational Autoencoder [21] (VAE) to reduce computational complexity, becoming the mainstream solution. Early video diffusion models [2, 3, 12, 13] were mostly built upon pretrained image diffusion models (typically U-Net [38] architectures), incorporating temporal modules to handle temporal consistency. However, their isolated processing of spatial and temporal information inherently limited their quality and consistency. With the release of Sora [4] and its epoch-making generation quality, researchers notice the potential of Diffusion Transformer [31] (DiT) for video generation. Recent DiT-based methods [9, 16, 22, 48, 53] apply a unified manner to model the video in spatial-temporal domain, and scale the capability of DiT by more parameters, data and computing resources, achieving more high-quality and consistent video generation.

2.2. Image Stylization

Gatys et al. [10] pioneered image stylization using neural networks. Early methods [6, 10, 18, 36] relied on the statistical descriptors (such as gram matrices, mean and standard deviation, and histograms) extracted from a pretrained VGG [42] network to represent and transfer style infor-

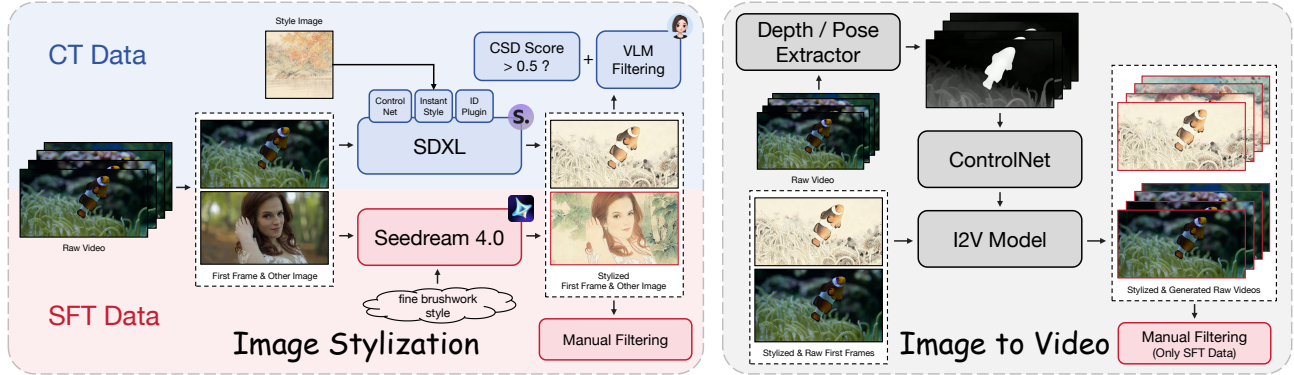


Figure 2. Data Curation Pipeline. We propose generating the training data with two key steps: image stylization followed by image to video. Considering the characteristics of different image stylization techniques, we construct a CT dataset and a SFT dataset, where SDXL (equipped with ControlNet, InstantStyle, and ID plugin) and Seedream 4.0 are selected as their stylization models, respectively. For image to video, we utilize ControlNets to enhance the motion consistency between the generated stylized and raw videos. To ensure the data quality, we additionally apply automatic filtering for CT data and manual filtering for SFT data.

mation. However, due to the limitations of the capability of generation and style extraction, they could only achieve simple texture and color transfer, often resulting in suboptimal visual quality. Recent methods [34, 49, 52, 54] benefit from the advances of diffusion models to improve basic quality, and CLIP [35] to extract high-level semantic information from style images. StyleTokenizer [24] further improves the style extractor with contrastive learning using a self-collected style dataset. Given the importance of high-quality datasets for stylization, OmniStyle [51] constructs a large-scale paired dataset using six state-of-the-art (SOTA) image stylization methods, and leverages a unified DiT backbone to extract style features and generate images, yielding new SOTA performance.

2.3. Video Stylization

Extending image-based tasks to video domain is a major trend in current research, and stylization is no exception. TokenFlow [11] and AnyV2V [23] achieve video stylization by leveraging image stylization techniques to stylize the first frame or key frames and then propagate it to the entire video sequence. However, these approaches can not perform video stylization independently, and rely on a time-consuming DDIM [45] inversion. UniVST [46] further DDIM inverts the style image and leverage AdaIN [18] to guide the denoising progress of noisy video by the inverted features of style. StyleCrafter [26] utilizes CLIP to extract style features and inject these features into the denoising U-Net via dual cross-attention. More recently, StyleMaster [56] upgrades to DiT backbone and incorporates both global and local style extractors, resorting to StillMoving [5] to train a LoRA [17] for temporal attention to bridge the gap between image and video. However, this scheme requires explicit temporal modeling within the base model,

which deviates from mainstream architectures. Moreover, a limitation shared by all these methods is their lack of stylized video datasets, resulting in suboptimal visual quality and temporal consistency.

3. Method

3.1. Data Curation Pipeline

Given the fact that current image generation / editing models are superior to the video counterpart in terms of visual quality, structure, aesthetics and text following, we propose generating stylized video datasets with two key steps: (1) leverage the SOTA image stylization models to stylize the first frame of raw video; (2) utilize the I2V model to generate stylized video from the stylized first frame. Our data curation pipeline is illustrated in Fig. 2. It is noteworthy that a high-quality first frame serves as crucial cues (e.g., style constraints and content anchors) to improve the overall quality of the entire video generated by I2V model.

To obtain the high-fidelity stylized first frame, we select InstantStyle [49] and Seedream 4.0 [41] as our image stylization models, which are proficient in style-image-guided and text-guided stylization, respectively. InstantStyle is a SDXL [33] plugin, which we further equip with a depth ControlNet [58] and ID plugin [14] to constrain the consistency of structure and face identity. It is worthy noting that the text-guided stylization model typically produces better visual quality and style consistency, while the style-image-guided stylization model allows us to generate images with greater style diversity. Thus, we construct two datasets: (1) a large-scale stylized dataset for Continual Training (CT) generated via InstantStyle to ensure the core video stylization capability and generalization of DreamStyle; (2) a small-scale higher-quality stylized dataset for Supervised

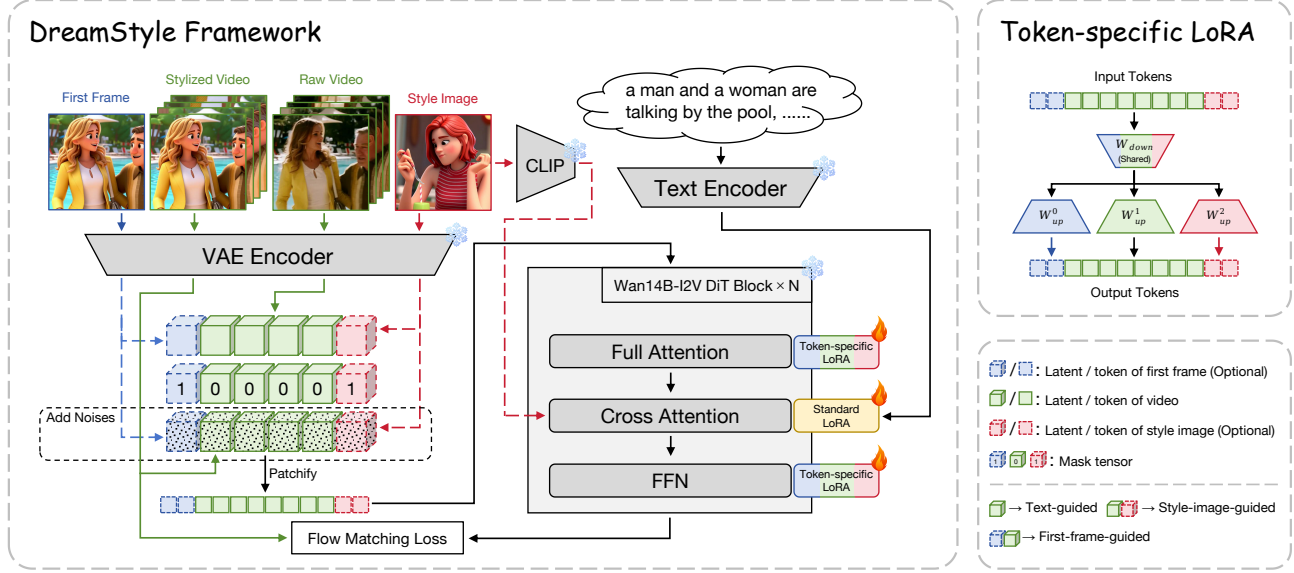


Figure 3. Overview of DreamStyle Framework. DreamStyle is built on the Wan14B-I2V model, integrating the text and raw-video conditions through the cross-attention and image channels of the base model, while the first-frame and style-image conditions serve as additional frames concatenated to the start and end of the frame sequence. We train it using a standard flow matching loss and a token-specific LoRA that contributes to distinguishing different condition tokens.

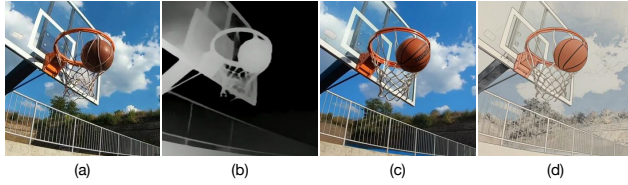


Figure 4. Example that depth fails to capture accurate detail. (a) The raw video frame, (b) the extracted depth map, (c) the generated realistic frame, (d) the generated stylized frame.

Fine-Tuning (SFT) generated with Seedream 4.0 to elevate the upper bound of DreamStyle.

It is critical to ensure the motion consistency between stylized video and raw video, so that we are able to construct stylized-raw video pairs. To this end, we customize two ControlNets (with control conditions of depth and human pose, respectively) for our in-house I2V model. The depth ControlNet is well-suited for general cases, while the human pose ControlNet offers a more precise control of human motion and especially allows for a larger deformation of driven objects without losing motion coherence. As illustrated in Fig. 4, we observe that directly animating the stylized first frame using the control conditions extracted from the raw video and then making paired data proves to be suboptimal. This is because neither depth nor pose can fully capture the complex motion dynamics of the raw video, ultimately resulting in motion mismatches between stylized and raw videos. Thus, we adjust to utilizing the same con-

trol condition to drive the generation of both stylized and raw video frames, aiming to mitigate such mismatches.

We formally denote our dataset as:

$$\mathcal{D} = \{(\mathbf{x}_i^{raw}, \mathbf{x}_i^{sty}, \mathbf{t}_i^{ns}, \mathbf{t}_i^{sty}, \mathbf{s}_i^{1...K}) | i = 1, 2, \dots, N\} \quad (1)$$

where $\mathbf{x}_i^{raw}$ and $\mathbf{x}_i^{sty}$ are the raw and stylized videos, $\mathbf{t}_i^{ns}$ and $\mathbf{t}_i^{sty}$ are the text prompts that exclude / include style descriptions, and $\mathbf{s}_i^{1...K}$ denotes K style reference images. To obtain $\mathbf{t}_i^{ns}$ and $\mathbf{t}_i^{sty}$, we utilize a Visual-Language Model [57] (VLM) to parse the stylized video $\mathbf{x}_i^{sty}$ and then generate the corresponding video caption. We restrict the VLM to exclude any style-related attributes (e.g., artistic genre, color palette, texture pattern, and hue) when generating $\mathbf{t}_i^{ns}$, so that $\mathbf{t}_i^{ns}$ contains only style-irrelevant descriptions. Regarding $\mathbf{s}_i^{1...K}$, we stylize K additional images using the same guided condition (text prompt or style reference) as $\mathbf{x}_i^{sty}$. For the CT dataset, we further filter out those $\mathbf{s}_i^{1...K}$ with low style consistency detected by VLM and CSD [44] score, while we opt for manual filtering for the SFT dataset. Additionally, we manually verify the content consistency between each raw video and its stylized video in the SFT dataset. Finally, such two datasets enable our DreamStyle to support all three video stylization tasks.

3.2. DreamStyle Framework

As shown in Fig. 3, our DreamStyle framework is built upon the Wan14B-I2V [48] base model that incorporates additional image condition channels before the patchify layer. This design allows for the injection of raw video condition

via these channels, rather than the in-context frames injection adopted in UNIC [55]. A major advantage is that it involves minimal extra computational overhead, ensuring DreamStyle retains the efficiency of the original I2V model. Overall, we have four types of conditions to inject into the I2V model: the raw video along with three guided conditions of style, which are described in detail as follows.

(1) Text condition. We reuse the original textual cross-attention layers of Wan14B-I2V without introducing modifications. **(2) First-frame condition.** We feed the stylized first frame into the original image condition channels and set the mask channels of the first frame to 1.0 in the same manner as the base model. **(3) Style-image condition.** Assuming that $\mathbf{z}^s \in \mathbb{R}^{C \times 1 \times H \times W}$ (omit the subscript i) is the VAE encoded latent of the style reference image $\mathbf{s}_i^j$, we construct the final I2V model’s input tensor for the style image via channel-wise concatenation:

$$\mathbf{z}_t^s = \text{add_noise}(\mathbf{z}^s, t) \oplus_c \mathbf{1}_{4 \times 1 \times H \times W} \oplus_c \mathbf{z}^s \quad (2)$$

where $\oplus_c$ denotes the concatenation operation along the channel dimension, $\text{add_noise}(\cdot, t)$ is the noise injection function of flow matching [25] at timestep $t \in [0, 1]$. The number of mask channels is 4 in Wan14B-I2V, thus $\mathbf{1}_{4 \times 1 \times H \times W}$ represents a mask tensor filled with a constant value 1.0. Wan14B-I2V also includes a native CLIP image feature branch, which we employ to inject high-level semantic features of the style reference image, thereby enhancing the consistency of style-related semantic information. **(4) Raw-video condition.** We first encode the raw video $\mathbf{x}_t^{\text{raw}}$ and stylized video $\mathbf{x}_t^{\text{sty}}$ to obtain their latents $\mathbf{z}^{\text{raw}}, \mathbf{z}^{\text{sty}} \in \mathbb{R}^{C \times F \times H \times W}$, and then channel-wise concatenate them with mask tensor $\mathbf{0}_{4 \times F \times H \times W}$:

$$\mathbf{z}_t^v = \text{add_noise}(\mathbf{z}^{\text{sty}}, t) \oplus_c \mathbf{0}_{4 \times F \times H \times W} \oplus_c \mathbf{z}^{\text{raw}}. \quad (3)$$

Here we adopt the mask value 0.0, following the principle of minimal modification of the base model. For the style-image-guided mode, $\mathbf{z}_t^s$ is treated as an additional frame and concatenated to the end of $\mathbf{z}_t^v$ via frame-wise concatenation $\oplus_f$: $\mathbf{z}_t^v \oplus_f \mathbf{z}_t^s$, enabling the model to incorporate the style-image condition. Similarly, for the first-frame-guided mode, the first-frame tensor $\mathbf{z}_t^{1st}$ is concatenated to the beginning of $\mathbf{z}_t^v$ via frame-wise concatenation: $\mathbf{z}_t^{1st} \oplus_f \mathbf{z}_t^v$.

To retain the base model’s inherent generative capabilities, we adopt LoRA to train our DreamStyle. After patchification, the three conditions, first-frame, style-image, and raw-video, are transformed into their corresponding token sequences. However, these tokens serve distinct semantic roles, thus using a standard LoRA leads to inter-token confusion. Inspired by HydraLoRA [47], we propose adopting a modified LoRA with token-specific up matrices in full attention and feedforward (FFN) layers. That means, for an input token $\mathbf{x}_{in}$, we first project it using a shared down matrix $\mathbf{W}_{down}$, and then compute the output residual token

$\mathbf{x}_{out} = \mathbf{W}_{up}^i \mathbf{W}_{down} \mathbf{x}_{in}$ with a specific up matrix $\mathbf{W}_{up}^i$ according to the token type $i \in \{0, 1, 2\}$, which is analogous to a LoRA MoE [8] with manual routing. Such a LoRA enables the model to learn adaptive features tailored to the three types of tokens, and still be trained stably due to the large proportion of shared parameters.

3.3. Training

We follow the same optimization objective as flow matching to train our DreamStyle. Formally, we denote our model as $\mathbf{v}_\theta$, a function with five inputs: the video tensor $\mathbf{z}_t^v$, the timestep t , the first-frame tensor $\mathbf{z}_t^{1st}$, the style image tensor $\mathbf{z}_t^s$ and the text prompt $\mathbf{t}^{ns/sty}$. In each training batch, we randomly sample style conditions according to predefined ratios, thus the training objective is:

$$\mathcal{L}(\theta) = \begin{cases} \mathbb{E}_{\mathcal{D}} \|\mathbf{v}_\theta(\mathbf{z}_t^v, t, \emptyset, \emptyset, \mathbf{t}^{sty}) - (\mathbf{z}^{sty} - \epsilon)\|^2 & \text{(I)} \\ \mathbb{E}_{\mathcal{D}} \|\mathbf{v}_\theta(\mathbf{z}_t^v, t, \emptyset, \mathbf{z}_t^s, \mathbf{t}^{ns}) - (\mathbf{z}^{sty} - \epsilon)\|^2 & \text{(II)} \\ \mathbb{E}_{\mathcal{D}} \|\mathbf{v}_\theta(\mathbf{z}_t^v, t, \mathbf{z}_t^{1st}, \emptyset, \mathbf{t}^{ns}) - (\mathbf{z}^{sty} - \epsilon)\|^2 & \text{(III)} \end{cases} \quad (4)$$

where $\epsilon \in \mathcal{N}(0, 1)$ is a Gaussian noise and (I) $\sim$ (III) correspond to the loss terms for the text-guided, the style-image-guided and the first-frame-guided tasks, respectively. As mentioned in Sec. 3.1, we make two datasets with different scales and quality, thus adopting a two-stage training strategy. In the first stage, we train DreamStyle on the large-scale CT dataset, allowing the model to learn diverse styles and establish a foundational capability to handle all three style conditions. In the second stage, a higher-quality SFT dataset is used to further finetune DreamStyle, aiming to improve visual quality and style consistency.

4. Experiments

4.1. Implementation Details

Through the data curation pipeline (Sec. 3.1), we construct 40K and 5K stylized-raw video pairs for the CT stage and SFT stage training, where the video resolution is 480P and the length is up to 81 frames. In the CT dataset, each sample includes exactly one style reference while the samples from the SFT dataset contain 1 $\sim$ 16 style images, with one randomly selected for training. During the training, we empirically set the sampling ratio of the three style conditions (text-guided, style-image-guided and first-frame-guided) to 1 : 2 : 1. The training process is performed on 8 NVIDIA A100 GPUs, with each GPU accommodating a per-GPU batch-size of 1. To stabilize training, we further adopt a 2-step gradient accumulation strategy, resulting in a larger effective batch size of 16. We train DreamStyle for 6,000 and 3,000 iterations in the CT and SFT stages, respectively, using a LoRA with a rank of 64 and AdamW [27] optimizer with a learning rate of 4×10^{-5} . The total training process requires about 1,700 A100 GPU hours.

Table 1. Quantitative comparison. The best and second best results are shown in **bold** and underline.

Condition	Method	Metrics						
		CLIP-T / CSD Score	DINO Score	Dynamic Degree	Image Quality	Aesthetic Quality	Subject Consistency	Background Consistency
Text	Luma	0.132	0.406	0.766	<u>0.739</u>	0.572	0.934	0.942
	Pixverse	<u>0.155</u>	0.451	0.766	0.746	<u>0.628</u>	<u>0.948</u>	<u>0.951</u>
	Runway	0.154	<u>0.504</u>	<u>0.809</u>	0.725	0.606	0.940	0.944
	DreamStyle	0.167	0.584	0.894	0.738	0.656	0.952	0.956
Style Image	StyleMaster (T2V)	0.198	-	0.289	0.723	0.610	0.936	0.935
	DreamStyle (T2V)	0.532	-	<u>0.689</u>	<u>0.722</u>	0.641	0.950	0.961
	DreamStyle (V2V)	<u>0.515</u>	0.526	0.867	0.704	<u>0.635</u>	<u>0.938</u>	<u>0.948</u>
First Frame	VACE	0.689	0.716	0.889	0.716	0.573	0.922	<u>0.930</u>
	VideoX-Fun	<u>0.766</u>	<u>0.702</u>	0.844	<u>0.726</u>	<u>0.594</u>	0.915	0.924
	DreamStyle	0.851	0.640	<u>0.856</u>	0.731	0.630	<u>0.919</u>	0.932

4.2. Settings

For text-guided video stylization, we curate 50 videos paired with style prompts crafted by a designer as our test set. Since no open-source models specialized in text-guided video stylization are currently available, we opt to compare DreamStyle against three commercial models: Luma [28], Pixverse [32] and Runway [39]. We further expand the aforementioned test set to 90 videos and 15 style images (each style image is randomly paired with 6 videos) to evaluate the style-image-guided task. As a baseline, we select StyleMaster [56], the only open-source DiT-based method that supports style-image-guided video stylization in an end-to-end manner. For first-frame-guided task, we reuse these 90 videos and generate stylized first frames for each video using image stylization methods, and then choose VACE [20] and VideoX-Fun [1] as our competitors.

To evaluate the style consistency, we utilize the CSD [44] score as the quantitative metric. Specifically, the CSD score is computed between the style reference image and each frame of the generated video for style-image-guided task, while for first-frame-guided task, we evaluate this metric between the stylized first frame and all subsequent frames. For text-guided stylization, we employ ViCLIP [50] to measure the similarity between user prompt and stylized video. Moreover, structure preservation is evaluated using the cosine similarity of the patch features (excluding the CLS token) extracted from DINOv2 [30]. We further assess the overall quality of stylized video with five metrics from VBench [19]: dynamic degree, image quality, aesthetic quality, subject consistency, and background consistency.

4.3. Comparisons

Quantitative Comparison. As shown in Table 1, we conduct a comprehensive comparison across three video stylization tasks. In text-guided video stylization, DreamStyle achieves the highest CLIP-T (we measure text-video simi-

larity using only the style prompts, thus the CLIP-T is overall lower) and DINO score, indicating that it outperforms the other methods in both style prompt following and structure preservation. This superiority is further evidenced in the visual results in Fig. 5. For the overall video quality assessment, our method also has advantages in most metrics except image quality. Notably, image quality exhibits a negative correlation with dynamic degree, since a high dynamic video tends to involve motion blur, thereby decreasing this metric. Due to the incomplete open-source of StyleMaster, it supports only T2V instead of V2V stylization. Thus, we include an additional result for DreamStyle in T2V mode, where the video condition is set to empty. Quantitative metrics demonstrate the superior performance of our method, particularly in the aspects of style consistency and dynamic degree. Despite not being explicitly trained for T2V, DreamStyle naturally inherits this capability from the base model thanks to the LoRA training and outperforms its V2V counterpart in most metrics due to fewer constraints. In first-frame-guided video stylization, DreamStyle presents the optimal style consistency (CSD score) and either the best or second-best video quality metrics. Since the stylized first frames (especially those with geometric deformation) occasionally conflict with the structure of the input video, our method, despite its superior style consistency, is inferior to VACE and VideoX-Fun in structure preservation (DINO score). However, the visual results in Fig. 5 confirm that it can still maintain the primary structural elements of the input video.

Qualitative Comparison. Fig. 5 presents the visual comparisons between our DreamStyle and the competitors. In text-guided video stylization, Luma tends to generate videos with dark tones, and the subject pose, color, and content of its results deviate far from the input videos. Pixverse achieves a higher pose consistency but still suffers from content distortion (e.g., the camera in the left case and the bowknot in the right case). Runway often produces videos

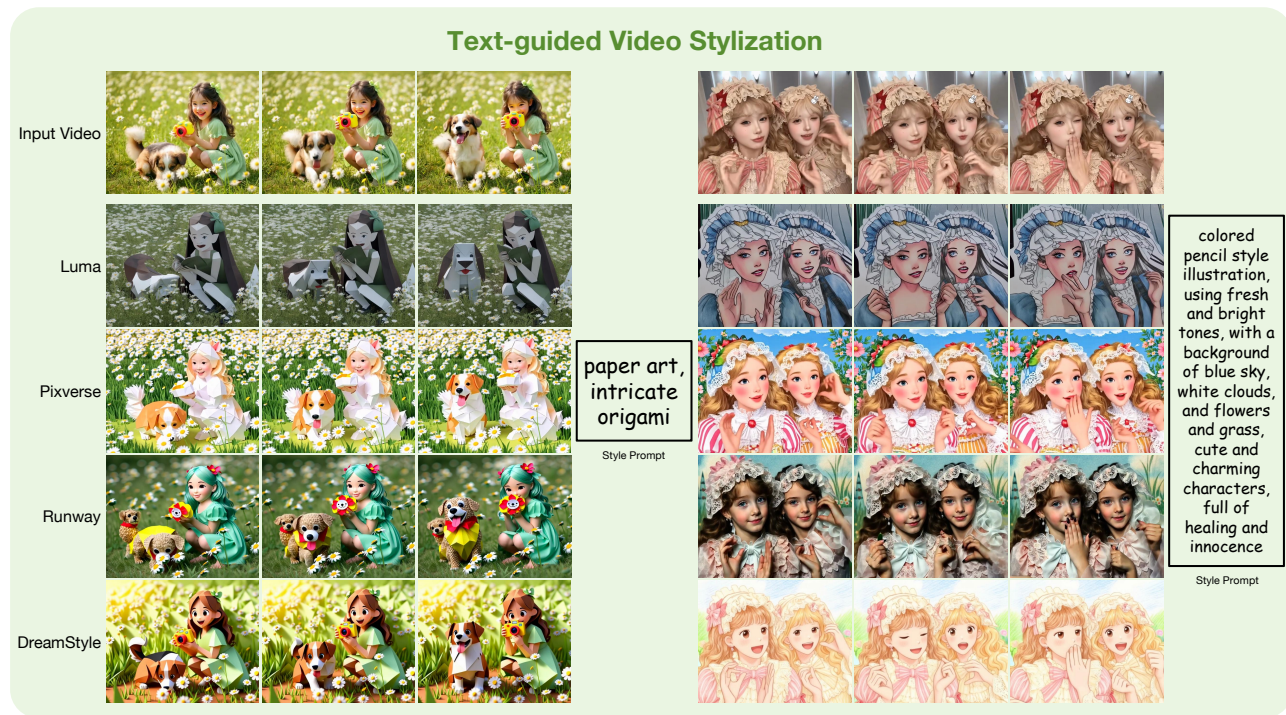


Figure 5. Qualitative comparison on three video stylization tasks.

with a realistic style bias, failing to accurately render the correct style. By contrast, our DreamStyle not only follows the style prompt but also achieves superior consistency with the input videos in terms of subject pose, color, and content. In style-image-guided video stylization, StyleMaster exhibits limited capability in simple color and texture transfer, while our method can further handle the styles involving geometric shapes. In first-frame-guided video stylization, both VACE and VideoX-Fun struggle to preserve the stylized first frame in the left case. For the right case, although they are able to maintain the major style of the given first frame, serious style degradation occurs in the subsequent frames. By comparison, DreamStyle demonstrates higher style consistency across the stylized first frame, the generated first frame, and all subsequent frames.

Table 2. Quantitative comparison of ablation studies.

	w.o. Token-specific LoRA	Only CT Data	Only SFT Data	Full
CSD Score	0.413	0.459	0.535	0.515
DINO Score	0.518	0.547	0.483	<u>0.526</u>

4.4. Ablation Studies

Token-specific LoRA. The proposed token-specific LoRA plays a critical role in mitigating interference among distinct condition tokens. To validate this, we design an ablation experiment where DreamStyle is trained with a standard LoRA—here, different condition tokens are distinguished solely through frame positions and mask values (1 for first-frame tokens, 0 for video tokens, and -1 for style-image tokens). We focus on the style-image-guided stylization task for evaluation. As shown in Table 2, the standard LoRA exhibits a significantly negative influence on style consistency (CSD score) and slightly reduces structure preservation (DINO score). Visual evidence in Fig. 6 further indicates this point, where the problems of style degradation (first row) and style confusion (second row) arise in the absence of token-specific LoRA.

Datasets. To validate the necessity of both datasets (with distinct scales and quality) and two-stage training, we conduct an ablation experiment where DreamStyle is trained on only the CT dataset, only the SFT dataset, and both of them in two stages. Due to the limited quality and style consistency of the CT dataset, training exclusively on it yields suboptimal stylization performance. Both the quantitative metric (lower CSD score) in Table 2 and the visual result (failure to render the pixel pattern) in Fig. 7 confirm this shortcoming. Conversely, the SFT dataset contains manually filtered paired videos with high quality and strong style consistency, but its limited size makes it insufficient to adapt the I2V base model into a robust V2V model for stylization (particularly, there is no strict alignment between the paired videos due to the existence of geometric deformation).

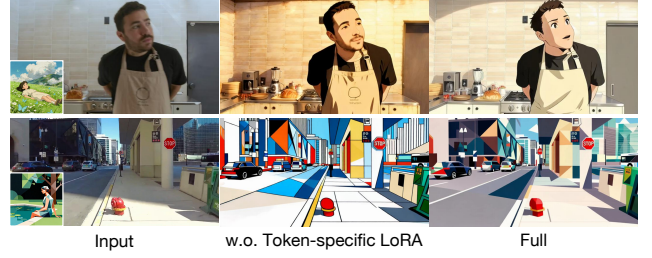


Figure 6. Impact of token-specific LoRA.



Figure 7. Visual comparison across different datasets.

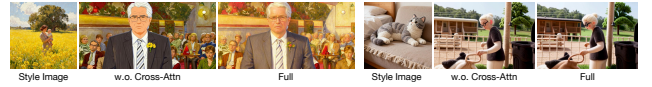


Figure 8. Impact of cross-attention injecting CLIP features.

tion). Thus, as shown in Table 2, training solely on it, despite achieving the best CSD score, exhibits the worst performance on structure preservation, which is also evidenced in Fig. 7 (the pose of the stylized panda differs from the input). As expected, training on the CT and SFT datasets in turn achieves a robust balance between style consistency and structure preservation.

Style Image Injection. We use the native cross-attention of Wan14B-I2V to inject the style image, whose CLIP features offer a global style cue that complements the VAE local features. We conduct training in the SFT stage without this injection, which as a result, slightly impairs the style consistency (visual comparison in Fig. 8) and reduces the CSD score from 0.515 to 0.484.

5. Conclusion

In this paper, we propose DreamStyle, the first unified framework for video stylization that supports three style conditions: text, style image, and first frame. We develop a systematic data curation pipeline to generate and filter paired videos. DreamStyle is built upon an I2V model, which can be efficiently extended to the V2V model without involving too much extra computation overhead. Moreover, to address the inter-token confusion among different style conditions within a unified model, we introduce a novel token-specific LoRA module. With our high-quality dataset, well-designed model architecture, and two-stage training strategy, DreamStyle archives competitive performance on various video stylization tasks.

References

- [1] AIGC-Apps. A more flexible framework that can generate videos at any resolution and creates videos from images. <https://github.com/aigc-apps/VideoX-Fun>. 6
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2
- [3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2
- [4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 2
- [5] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024. 2, 3
- [6] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017. 2
- [7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2
- [8] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, 2024. 5
- [9] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 2
- [10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2
- [11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [12] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2
- [13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [14] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024. 3
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3
- [18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 3
- [19] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6
- [20] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17191–17202, 2025. 6
- [21] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings*, 2014. 2
- [22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2
- [23] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhui Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024. Reproducibility Certification. 3
- [24] Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 3
- [25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 5

- [26] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Ying Shan, and Yujiu Yang. Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. *ACM Transactions on Graphics (TOG)*, 43(6):1–10, 2024. 2, 3
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5
- [28] Luma. Luma ai. <https://lumalabs.ai>. 6
- [29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 6
- [31] 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. 2
- [32] Pixverse. Pixverse ai video generator. <https://app.pixverse.ai>. 6
- [33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3
- [34] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8693–8702, 2024. 3
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [36] Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017. 2
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2
- [39] Runway. Runway ai image and video generator. <https://runwayml.com>. 6
- [40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2
- [41] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 3
- [42] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015. 2
- [43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 2
- [44] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 4, 6
- [45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3
- [46] Qianjian Song, Mingbao Lin, Wengyi Zhan, Shuicheng Yan, Liujuan Cao, and Rongrong Ji. Univst: A unified framework for training-free localized video style transfer. *arXiv preprint arXiv:2410.20084*, 2024. 3
- [47] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584, 2024. 5
- [48] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 4
- [49] 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. 3
- [50] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 6
- [51] Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer

- data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7847–7856, 2025. [3](#)
- [52] 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. [3](#)
- [53] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihang Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. [2](#)
- [54] Hu Ye, Jun Zhang, Sibol Liu, Xiao Han, and Wei Yang. Ip-adapt: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. [3](#)
- [55] Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. *arXiv preprint arXiv:2506.04216*, 2025. [2](#), [5](#)
- [56] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2630–2640, 2025. [2](#), [3](#), [6](#)
- [57] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. [4](#)
- [58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. [3](#)