

EchoStyle: Unlocking High-Fidelity Video Stylization with Reverse Data Synthesis

Huaqiu Li^{*1,2}, Jiahao Wang^{*2,3}, Sijia Cai^{†2}, Hualian Sheng², Bing Deng²,
Jieping Ye², and Wenhan Luo^{†1}

¹ The Hong Kong University of Science and Technology

² Alibaba Group

³ Xi'an Jiaotong University

<https://echostyle2026.github.io/>

lihuaqiu2025@gmail.com

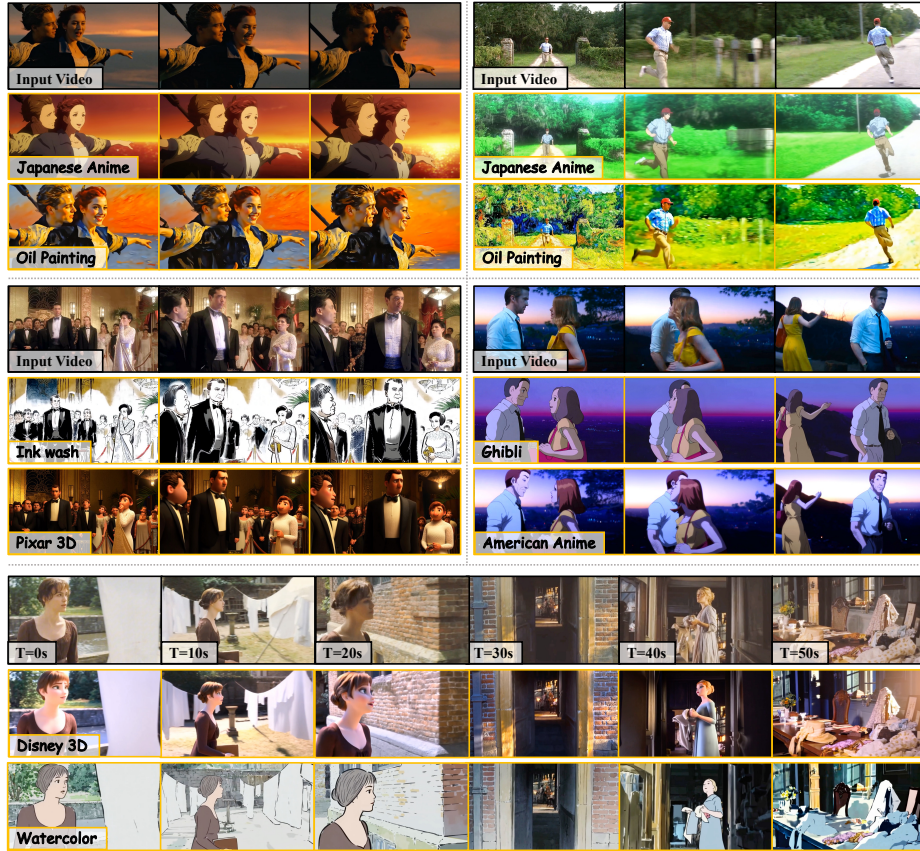


Fig. 1: EchoStyle provides a robust framework for video stylization across expansive artistic domains. By maintaining stringent temporal and motion coherence, it delivers visually compelling results that remain consistent even during long-video inference.

Abstract. While image stylization has been studied extensively, video stylization remains a critical and largely unsolved challenge in the field of intelligent content creation. Existing methods, usually utilizing a reference image as the style prior, suffer from content leakage, data scarcity and limited adaptability to long videos, leading to suboptimal results with severe style drift and motion distortion. For these issues, we present EchoStyle, a scalable text-driven framework to achieve high-quality stylization of videos with arbitrary lengths. To start with, we construct a video-to-video architecture to appropriately re-fuse the video content and the text style. To address data scarcity, we pioneer an automatic reverse-synthesis pipeline to establish V-Style20k, a large-scale stylization dataset of 20k high-quality video pairs. To facilitate long video stylization, we devise an init-follow-mode mechanism along with a sliding-window inference strategy. Extensive experiments demonstrate EchoStyle’s excellent performance across a wide range of artistic styles, even comparable to leading closed-source solutions.

Keywords: Video generation · Video stylization · Diffusion model

1 Introduction

Recent advancements in diffusion models have sparked remarkable progress in generative modeling, triggering the evolution of controllable image and video synthesis [11, 13, 18, 21, 36, 39, 43, 45, 47, 50]. Along with the development, video stylization, which aims to render a given video in a particular style, is garnering growing attention from both academia and industry for its wide-ranging applications such as artistic creation and advertisement production.

Previous works in video stylization predominantly focus on introducing a reference image as a style prior. Conventional methods [6, 10, 14, 30] leverage CNNs or ViTs as backbones, primarily prioritizing real-time efficiency. In recent years, there has been a significant shift towards adopting diffusion models as the foundational architecture. For instance, StyleCrafter [23] and StyleMaster [46] adapt pre-trained text-to-video diffusion models [2, 37] through post-training strategies. In parallel, training-free methods, such as AnyV2V [19] and Free-Vis [42], typically stylize keyframes via image editing models and then propagate the style through feature injection to ensure temporal consistency. Furthermore, frameworks like UniVST [32] have extended training-free capabilities to support localized video stylization.

However, these existing paradigms suffer from several critical limitations: **Content-style entanglement.** Since the reference image contains redundant style-irrelevant information besides style, the reliance on it often leads to unintended content leakage into the stylized videos [20, 44]. Consequently, commercial closed-source video models [8, 35] and several image stylization methods [20, 25, 33] concentrate on the text-driven roadmap to achieve concept-level

⁰ *: Co-first Author.

¹ †: Corresponding Author.

stylization and inherently avoid content leakage. Yet, despite their advancements, a reliable open-source text-driven video stylization paradigm remains elusive. **Video data scarcity.** The training of video stylization requires high-quality video pairs with the same content but different styles, which are extremely scarce. Existing video stylization methods either bypass training by handcrafted components [19, 32, 42] or conduct training on image datasets [3, 23, 46], both of which lead to severe style degradation and instability, especially under strong video dynamics. **Limited ability of extension.** Previous works are confined to short sequences (≤ 5 s) and fail to provide robust, scalable designs for long-duration generation. Therefore, developing a stable and expandable framework for text-driven video stylization remains a valuable and challenging problem.

To address the aforementioned challenges, we present EchoStyle, a text-driven framework designed for high-fidelity stylization of videos with arbitrary lengths. We formalize the text-driven video stylization as video-to-video generation and propose a streamlined yet effective architecture based on the Wan2.2-I2V [37] foundation to appropriately re-fuse the video content and the text style. To bridge the data gap caused by scarcity and inadequate quality, we devise an efficient and reusable data generation pipeline that employs reverse-synthesis to derive reference input videos from open-source stylish datasets. With such a curated pipeline, we construct V-Style20k, a pioneering large-scale video stylization dataset of 20k high-quality video pairs to drive the training process. To further achieve consistent stylization for long videos, we design a temporal recurrent generation approach, which highlights an init-follow-mode mechanism during training and a sliding-window strategy during inference. As shown in Fig. 1, extensive experiments demonstrate that our method achieves robust and stable stylization for both short and long videos, even in cases of complex content and intense dynamics.

We summarize our key contributions as follows:

- We propose a novel video-to-video framework tailored for text-driven stylization of videos with arbitrary lengths, which is naturally scalable owing to its simplicity.
- We pioneer a fully automated data curation pipeline to construct an innovative large-scale video stylization dataset, V-Style20k, which contains 20k high-quality video pairs.
- We incorporate an intricate init-follow-mode mechanism into the training along with a sliding-window inference strategy, enabling stable stylization for long videos spanning several minutes.
- Comprehensive experiments demonstrate that EchoStyle achieves superior dynamic style consistency as well as fine-grained content preservation in video stylization, yielding performance comparable to closed-source models.

2 Related Work

2.1 Image & Video Stylization

Stylization aims at applying the artistic style from the reference condition to either a given content image or video. For image stylization, early research, such as [6, 14], utilizes a pre-trained network as a style encoder and a carefully designed injection module. More recently, the advance of powerful text-to-image foundation models, such as Stable Diffusion [29] and FLUX [22], has driven a series of methods [28, 44] based on diffusion models. Building upon this, works like InstantStyle [38], StyleShot [7], and B-LoRA [5] have proposed more granular, time-aware and layer-aware injection strategies to better disentangle the infusion of style and content features.

In the field of video stylization, conventional methods [6, 10, 14] predominantly rely on CNN or ViT backbones, prioritizing low-latency inference and real-time efficiency. However, recent years have witnessed a paradigm shift toward diffusion-based architectures. For instance, StyleCrafter [23], StyleMaster [46] and PickStyle [26] adapt pre-trained T2V models through specialized post-training strategies, while Telestyle [49] tries to adapt pre-trained image generation models to the video domain. Concurrently, training-free pipelines have gained attention by leveraging non-end-to-end workflows. Approaches like AnyV2V [19] and FreeVis [42] typically stylize representative keyframes via image-editing models and subsequently propagate these stylistic attributes through feature injection to ensure temporal consistency. Furthermore, frameworks such as UniVST [32] have extended these training-free capabilities to support localized or region-aware video stylization.

2.2 Video-to-Video Frameworks

Existing research has extensively investigated Video-to-Video (V2V) translation frameworks. In particular, Video-P2P [24] facilitates real-world video editing through cross-attention manipulation, while TokenFlow [9] and AnyV2V [19] maintain temporal consistency through the propagation of cross-frame attention features. Other task-specific approaches, including Animate Anyone [13] and MotionCtrl [40], leverage Denoising UNet with temporal attention and Latent Video Diffusion Model [2] backbones. Furthermore, VACE [16] adopts a ControlNet-like mechanism to inject video guidance, achieving unified V2V generation. Despite these advancements, prevailing methodologies often emphasize generative diversity rather than input-output semantic alignment and face significant challenges in scaling to long videos. This underscores the necessity of a dedicated V2V framework specifically designed for video stylization.

3 Method

To facilitate high-quality and expandable text-driven video stylization, we introduce a novel video-to-video architecture to appropriately re-fuse the video

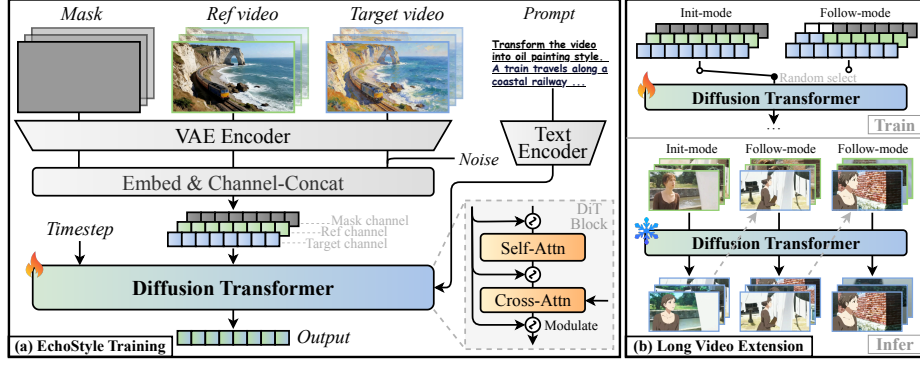


Fig. 2: Overview of the EchoStyle framework. (a) Training: We map the mask, reference video, and target video into the shared latent space, followed by visual alignment. These multi-channel embeddings are then fed into a DiT for stylized video generation, guided by textual prompts. (b) Long Video Extension: We propose an init-follow-mode training strategy. During training, these modes are randomly selected. In the inference stage, long videos are generated autoregressively, where init-mode generates the starting sequence and subsequent segments are produced via follow-mode.

content and the text style in Sec. 3.1. To address data scarcity, we present a robust data curation roadmap that utilizes a reverse-synthesis pipeline to construct a large-scale paired video dataset in Sec. 3.2. Finally, in Sec. 3.3, we detail an intricate init-follow-mode mechanism along with a sliding-window inference strategy for long-video stylization.

3.1 Constructing Text-Driven Video-to-Video Framework

Formally, we characterize text-driven video-to-video stylization through a set of input conditions: a text prompt p , which encompasses both target style and content descriptions, a reference video $\mathbf{R}$ to be stylized, and a mask $\mathbf{M}$ to define which frames to be modified. We expect the model to produce the output video $\mathbf{O}$ through the mapping:

$$\mathbf{O} = \Phi(\mathbf{R}, \mathbf{M}, p), \quad (1)$$

where $\Phi(\cdot)$ denotes the generative framework, and $\mathbf{M}$ is set to 0 for every frame.

The task of video stylization primarily hinges on addressing two core challenges: achieving rigorous content alignment with the reference video and developing expandable potential. To overcome these bottlenecks, EchoStyle adopts a streamlined yet effective architecture as illustrated in Fig. 2(a). To specify, we use the Variational Autoencoder (VAE) [17] to encode the reference video $\mathbf{R}$, the target video $\mathbf{V}$ into the latent space, yielding their respective latent representations, $\mathbf{z}_R$ and $\mathbf{z}_V$. Next, a timestep t is uniformly sampled from $1, \dots, T$. We then corrupt the target latent $\mathbf{z}_V$ with noise according to the diffusion process forward kernel, resulting in the noisy latent $\mathbf{z}_V^{(t)}$. Concurrently, $\mathbf{M}$, which shares the same spatial and temporal dimensions as the input videos, is then reshaped

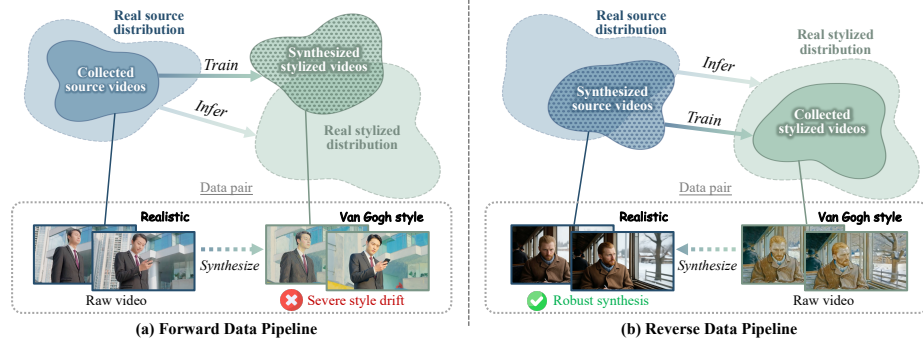


Fig. 3: Forward vs. Reverse Data Pipelines. (a) The forward pipeline suffers from significant distribution mismatch and style drift. (b) In contrast, our reverse pipeline leverages real stylized data as the source, achieving superior distribution alignment and more robust generation results.

to match the shape of the latent $\mathbf{z}_V^{(t)}$ and $\mathbf{z}_R$, yielding $\mathbf{z}_M$. Finally, we concatenate these three latent tensors along the channel dimension to form the unified model input $\mathbf{z}_{in}$ as:

$$\mathbf{z}_{in} = \text{Concat}(\mathbf{z}_R, \mathbf{z}_V^{(t)}, \mathbf{z}_M). \quad (2)$$

By introducing this unified and standardized visual alignment strategy, we consolidate diverse references $(\mathbf{R}, \mathbf{V}, \mathbf{M})$ into a unified latent. This approach ensures high semantic coherence, spatial and temporal alignments between the input video condition and the target $\mathbf{V}$.

This concatenated tensor is then fed into the backbone for noise prediction, which is constructed by stacking a series of Diffusion Transformer (DiT) blocks [27]. The conditioning information, including the timestep t and textual prompt p , is injected into the network through Adaptive Layer Normalization (adaLN) and Cross-Attention. The forward propagation of the DiT model can be denoted as:

$$\frac{d\mathbf{z}_V^{(t)}}{dt} = \epsilon_\theta = \text{DiT}(\mathbf{z}_{in}, t, p), \quad (3)$$

where ϵ_θ is the prediction from our model θ . We utilize the standard flow-matching loss as the training objective, which is to minimize the MSE between the prediction ϵ_θ and the ground-truth $(\mathbf{z}_V^{(0)} - \epsilon)$, then update the entire model in a LoRA [12] manner. The training objective can be denoted as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1]} \left\| \frac{d\mathbf{z}_V^{(t)}}{dt} - (\mathbf{z}_V^{(0)} - \epsilon) \right\|_2^2. \quad (4)$$

3.2 Curating High-Quality Video Pairs

Given the scarcity of large-scale, paired video stylization datasets, synthesizing data with existing models is a common strategy. However, AI-synthesized styl-

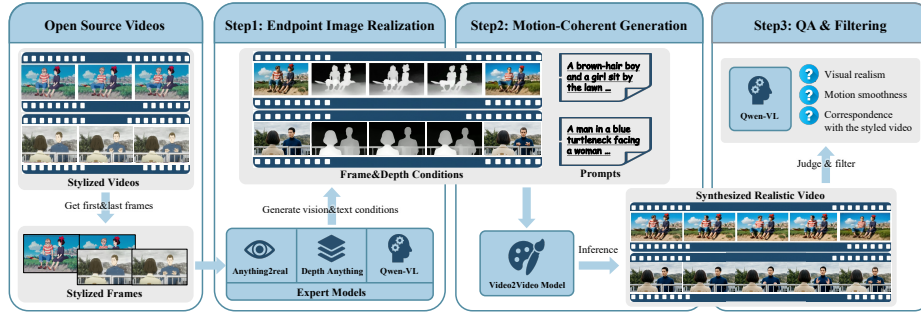


Fig. 4: Our data synthesis pipeline is organized into three stages: first, endpoint image stylization establishes the style for boundary frames; second, motion-coherent V2V generation produces semantically-aligned video sequences; and finally, automated filtering via VLMs is employed to vet video quality and discard suboptimal samples.

ized videos suffer from limited stylistic diversity, flickering, and style drift. As Fig. 3 illustrates, using such data as a training target would contaminate the objective distribution with these flaws, thereby capping the model’s potential. This limitation stems from the data bias inherent in current video2video tools, where the training sets are dominated by real-world imagery.

We effectively address the aforementioned challenges by inverting the direction of this synthesis process. To specify, we collect authentic stylized videos and synthesize their corresponding realistic counterparts to serve as source inputs. The primary advantage of this approach lies in defining authentic stylized videos as the target distribution, which ensures the purity of the learning objective and effectively circumvents the performance bottlenecks typically imposed by synthetic artifacts or stylistic drift. Furthermore, by strategically exploiting the intrinsic realism bias of pre-trained models, we achieve superior fidelity and stability in reference video construction during reverse synthesis, utilizing model bias as an advantage for high-quality data curation.

As shown in Fig. 4, our automated data collection pipeline proceeds as follows:

- **Stylized Video Collection.** We collect stylized video clips from open-source datasets and extract their initial and final frames.
- **Image Realization.** We translate the first and final frames of the stylized videos into realistic endpoint anchors using Qwen-Image [41] as the base model and Anything-to-Real LoRA for image-to-image translation.
- **Motion-coherent V2V Generation.** These transformed realistic images, combined with depth priors extracted from the original stylized clip, serve as conditions for VACE [16]. This depth and first & last frame-guided process ensures the generated realistic video maintains precise content coherence with the stylized video.
- **Quality Assessment.** To ensure data fidelity, we utilize Qwen-VL [1] to screen video pairs based on three key metrics: (i) generated visual realism,

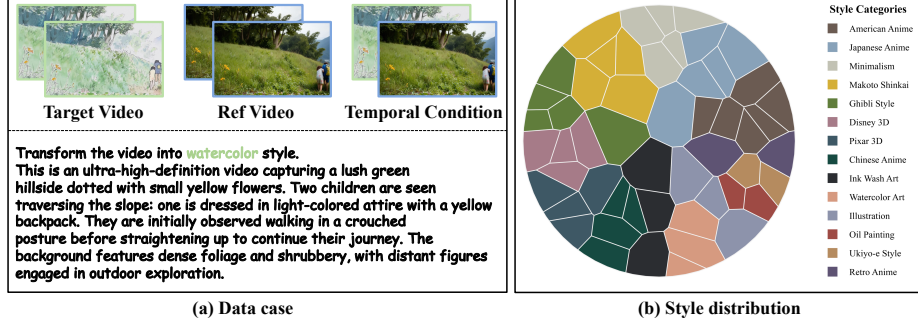


Fig. 5: Representative samples and statistical characterization of V-Style20k.

(ii) motion plausibility, and (iii) fine-grained temporal correspondence. Only pairs that meet these rigorous criteria are incorporated into the final training dataset.

***V-Style20k* Dataset.** Leveraging this reverse data pipeline, we establish *V-Style20k*, a large-scale paired video dataset designed to facilitate research in video stylization. The dataset comprises over 20,000 high-quality video pairs at a resolution of 480×832 , with each clip restricted to a duration of under 5 seconds. *V-Style20k* encompasses 14 diverse styles prevalent in cinematography, animation, and fine arts. Furthermore, to support text-driven tasks, each sample is meticulously annotated with descriptive captions using the Qwen-VL [1] vision-language model. We formally denote our dataset D as:

$$D = \{(p_j, \mathbf{R}_j, \mathbf{V}_j, \mathbf{T}_j) | j = 1, 2, \dots, N\}. \quad (5)$$

Specifically, each data sample comprises a text prompt p describing the intended style and content, an input reference video $\mathbf{R}$ serving as the content prior, a target stylized video $\mathbf{V}$, and a $\mathbf{T}$ serving as a temporal condition which will be thoroughly discussed in Sec. 3.3. We provide the illustration and statistics of our dataset in Fig. 5.

3.3 Expanding to Long-Video Stylization

To mitigate the pervasive memory bottlenecks in long-video generation and achieve expandable stylization, we propose a sliding-window-based generation framework powered by an init-follow-mode mechanism during training, as shown in Fig. 2(b). Specifically, in the follow mode, we introduce a temporal reference video $\mathbf{T}$, which is constructed by replacing the initial x -frame window of $\mathbf{R}$ with the corresponding stylized frames from $\mathbf{V}$. During training, we utilize VAE to compress $\mathbf{T}$ into $\mathbf{z}_T$. Then we replace $\mathbf{z}_R$ in Eq. (2) with $\mathbf{z}_T$ to realize dual-mode training, and modify the respective frames in the mask $\mathbf{M}$ as 1 to indicate that these frames need to be preserved, which can be denoted as:

$$\mathbf{z}_{\text{in}} = \text{Concat}(\mathbf{z}_T, \mathbf{z}_V^{(t)}, \mathbf{z}_M). \quad (6)$$

In each training iteration, we choose $\mathbf{z}_{\text{in}}$ from two modes with a fixed probability, ensuring that the network effectively balances its capacity.

During inference, instead of relying solely on the first frame, we utilize clip-guided conditioning to reduce the sensitivity to one single frame’s quality and enhance the temporal stability of the stylized output. We partition the long-form video into a sequence of shorter segments $\{\mathbf{R}_0, \mathbf{R}_1, \dots, \mathbf{R}_n\}$, where each consecutive pair $(\mathbf{R}_i, \mathbf{R}_{i+1})$ shares a temporal overlap of x frames. The initial segment $\mathbf{R}_0$ is processed following the short-duration pipeline. For subsequent segments $\mathbf{R}_i$ ($i > 0$), the generation is conditioned on the overlapping stylized frames from the previous output $\mathbf{O}_{i-1}$. Formally, let $\mathbf{O}_{i-1}[-x:]$ denote the last x frames of the preceding stylized segment. $\mathbf{M}_x$ means setting the first x frames of $\mathbf{M}$ to 1 to indicate that these frames should be modified. The inference for $\mathbf{R}_i$ is formulated as:

$$\mathbf{O}_i = \Phi(\mathbf{R}_i, \mathbf{M}_x, p_i, \mathbf{O}_{i-1}[-x:]), \quad (7)$$

This process ensures smooth and temporally coherent generation across long video sequences. The equation can be formulated as:

$$\mathbf{O}_{\text{long}} = \mathbf{O}_0 \parallel \sum_{i=1}^n \mathbf{O}_i[x:]. \quad (8)$$

where $\parallel$ denotes the concatenation along the temporal axis. This init-follow-mode helps us to seamlessly integrate the training and inference processes in variable-length video stylization into a unified pattern, not only enhancing the structural simplicity but also significantly reducing computational costs.

4 Experiment

4.1 Implementation Details

Data Preparation. To facilitate our training process, we utilize *V-Style20k* during our training process. Videos are center-cropped to a spatial resolution of 480×480 to maintain consistency. We adopt a temporal bucket strategy to handle varying sequence lengths, with frame counts ranging from 33 to 69 frames, all sampled at a frame rate of 16 FPS.

Base Model and Components. Our framework is initialized with the pre-trained weights of Wan2.2-I2V-14B, spanning both high-noise and low-noise regimes. During the fine-tuning stage, we freeze the VAE and text encoders, updating only the LoRA parameters integrated into the DiT backbone. Specifically, we inject LoRA into all `nn.Linear` layers with a rank of $r = 64$ and $\alpha = 32$, utilizing Gaussian initialization. Training is orchestrated via torchrun on a cluster of 16 NVIDIA A100 GPUs. We set the global seed to 42 and apply a per-GPU batch size of 1, utilizing a *seed+rank* strategy to ensure multi-process decorrelation. The model is trained for 20,000 steps, totaling approximately 80 hours of computation.

Training Strategy. To enhance the model’s robustness and controllability, we employ a diverse sampling strategy for text and video conditioning during each iteration. For textual guidance, the model is provided with full captions (a concatenation of style instructions and content descriptions) with a probability of 0.8; otherwise, only style instructions are used. Regarding visual conditioning, we utilize reference video $\mathbf{R}$ with all-zero mask $\mathbf{M}$ in 80% of the samples. For the remaining 20%, we use the temporal reference video $\mathbf{T}$ with the modified mask $\mathbf{M}_x$, and x is set to 16 during training. We leverage mixed-precision training, where non-LoRA weights are converted to `bfloat16` to reduce memory overhead. Optimization is performed using the AdamW optimizer with $\beta = (0.9, 0.999)$, a weight decay of 0.01, and a constant learning rate of 2×10^{-5} .

4.2 Comparison Experiment

To benchmark our method against both open-source and closed-source models, we assemble an evaluation set of 50 video-instruction pairs featuring nine representative styles. Due to the scarcity of text-based open-source tools, we compare our model with leading commercial solutions (Runway, Kling-O1 [35], and Seedance 2.0 [8]) and open-source V2V baselines (VACE [16] and AnyV2V [19]). To bridge the gap of the open-source V2V models, which support frame-conditioned generation, we utilize Qwen-Image [41] to stylize the first frame as a reference, ensuring a consistent and fair evaluation across different modalities.

Quantitative evaluations are performed across three dimensions: style quality, video quality, and human preference. To evaluate style similarity, we leverage Gemini-3-pro [34] to perform automated pairwise comparisons. Specifically, for each test case, the model is tasked with identifying which generated video aligns more closely with the reference style; the performance of each method is then quantified by its overall win rate. Furthermore, we assess style consistency by calculating the CSD scores [31] between each subsequent frame and the initial frame. This metric serves to measure the stylistic stability across the temporal dimension, ensuring that the stylistic attributes remain coherent throughout the video sequence. Following the established settings of VBench [15], we assess video quality through motion smoothness and imaging quality. In addition to standard metrics, we employ Gemini-3-pro [34] to perform a comprehensive automated assessment of the results, focusing on dynamic quality, static quality, and quality. Furthermore, our human preference evaluation focuses on both artistic scores and fine-grained content alignment with the input videos. We refer the reader to the Supplementary Material for more details.

The qualitative and quantitative comparisons are summarized in Fig. 6, Tab. 1 and Tab. 2. In summary, EchoStyle demonstrates superiority over existing baselines in style quality and human preference. Regarding video quality, EchoStyle follows closely behind SOTA closed-source methods, Kling-O1 and Seedance 2.0. Significantly, we secure the top-tier ranking in dynamic quality and the second-highest score in aesthetic quality, highlighting the immense artistic utility and excellent temporal quality of our proposed method and offering a powerful alternative for professional-grade artistic video production.

Table 1: Quantitative comparison between EchoStyle and SOTA methods. The **best** and second-best results are highlighted in **bold** and underline, respectively.

Method	Style Quality		Video Quality				
	Style Similarity	Style Consis.	Motion Smooth.	Imaging Qual.	Static Qual.	Dynamic Qual.	Aesthetic Qual.
AnyV2V	0.351	<u>0.877</u>	0.957	0.654	0.733	0.611	0.815
VACE-Depth	0.280	0.805	0.972	0.687	0.845	0.700	0.837
VACE-Flow	0.394	0.819	0.960	0.681	0.856	0.732	0.852
Runway	0.239	0.860	<u>0.982</u>	0.704	0.721	0.637	0.747
Kling-O1	0.809	0.883	0.984	0.746	0.879	<u>0.775</u>	0.806
Seedance2.0	0.628	0.831	0.980	<u>0.725</u>	<u>0.860</u>	0.741	0.885
EchoStyle	<u>0.802</u>	0.895	0.966	0.696	0.850	0.783	<u>0.878</u>

Although Kling-O1 and Seedance 2.0 offer high visual quality, they exhibit several drawbacks in the task of fine-grained video stylization. As shown in the bottom two subfigures of Fig. 6, Kling-O1 exhibits limited capacity in synthesizing complex scenes, often resulting in sparse details. Its generated results exhibit monotonous color schemes and simplistic lighting hierarchies, leading to lower human-rated style similarity

compared to our method. Runway, while preserving alignment, suffers from sub-optimal visual quality and limited expressiveness, failing to faithfully transition to the target styles. Seedance 2.0 achieves the best results in terms of human preference for content preservation. However, as shown in the top two subfigures of Fig. 6, it often struggles to maintain *high-frequency stylistic details*, such as those in oil painting and ink art styles. Furthermore, we observe that its output is prone to *style decoupling* between the subject and the background, as illustrated in Fig. 6 bottom-left case. For first-frame-guided baselines, VACE experiences significant “style loss”. As the video progresses, the stylized features increasingly revert to their realistic counterparts. In the stylized output, the stylistic attributes are progressively lost, converging towards the original, non-stylized frames. Additionally, AnyV2V is constrained by its architecture to sequences under 16 frames. In summary, EchoStyle consistently outperforms competing methods across style quality, video aesthetics, and human preference evaluation.

Table 2: Comparison of Human Preference. The **best** and second-best results are highlighted in **bold** and underline, respectively.

method	Style Similarity	Style Consistency	Content Preservation
AnyV2V	0.490	0.768	0.790
VACE-Depth	0.489	0.779	0.850
VACE-Flow	0.472	0.765	0.855
Runway	0.386	0.849	0.844
Kling-O1	<u>0.815</u>	<u>0.853</u>	0.833
Seedance2.0	0.792	0.821	0.889
EchoStyle	0.826	0.870	<u>0.860</u>

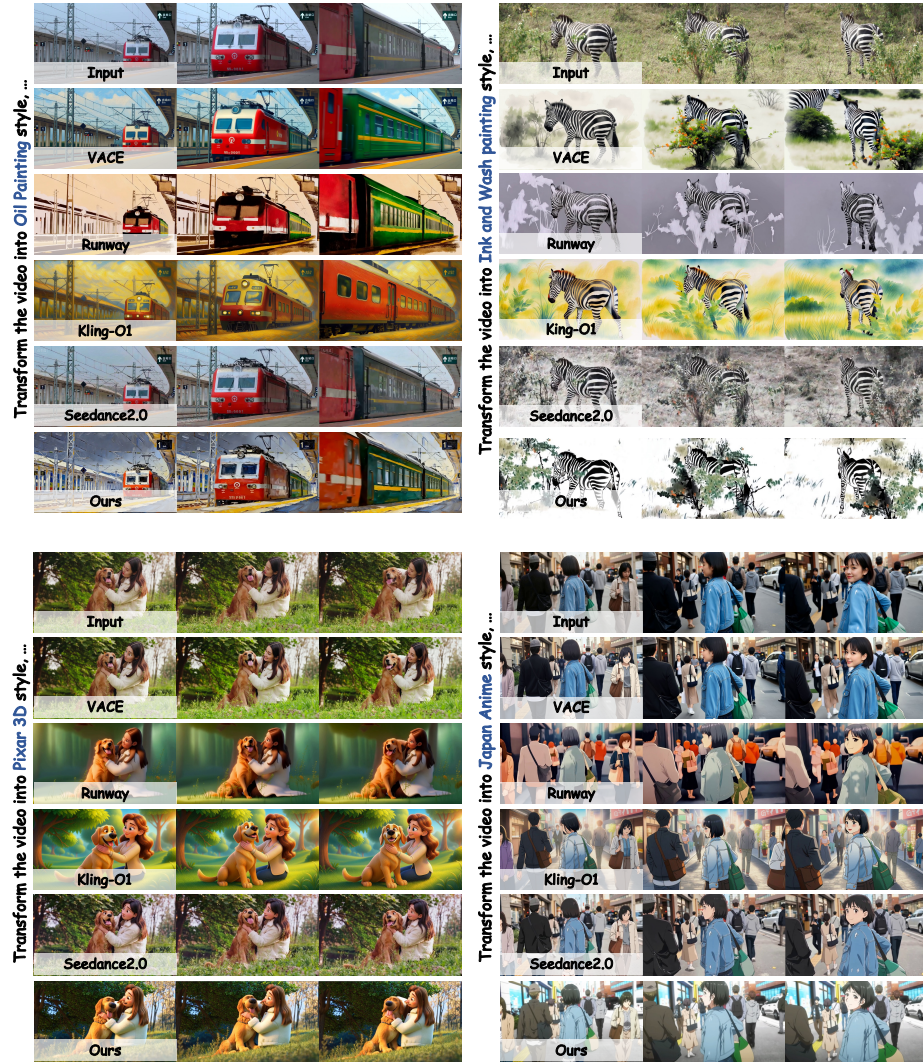


Fig. 6: Visual comparison of video stylization results. EchoStyle demonstrates excellent visual quality and stronger temporal coherence.

4.3 Long Video Extension

To address the inherent challenges of content preservation and quality degradation in long-duration video generation, we systematically evaluate the robustness of our framework. Our experiments utilize high-quality cinematic long shots characterized by complex spatial dynamics and intricate illumination variations, which impose stringent requirements on temporal stability.



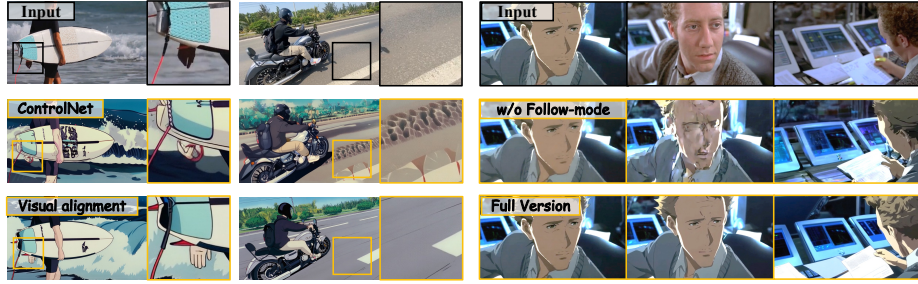
Fig. 7: Visual demonstration of temporally consistent stylization on extended video sequences. Our method demonstrates superior performance in maintaining content preservation, illumination depth, and artistic texture.

For comparison, we employ Kling-O1 as the baseline. Specifically, for its long-duration extension, we utilize the last frame of the preceding video segment in conjunction with the reference video as conditioning prompts. The qualitative results are presented in Fig. 7. Both models successfully achieve the target stylistic transformation, characterized by exaggerated facial features (e.g., enlarged eyes) and smoothed skin textures typical of modern 3D animation. However, EchoStyle demonstrates superior stylistic distillation and visual refinement, while Kling-O1 tends to produce overly simplified textures and flatter lighting. At the same time, Kling-O1 occasionally suffers from semantic drift, leading to quicker error accumulation, which causes the character motions and environmental scenes in the latter stages of the video to deviate significantly from the original source content.

4.4 Ablation Study

We conduct a series of ablation experiments to validate the effectiveness of the respective designs. Due to space constraints, we refer the reader to the Supplemental Material for further details and analysis.

Effectiveness of Model Framework. We compare our utilized *Visual Alignment* strategy against two common baselines: *ControlNet* and *In-context conditioning*. As Fig. 8(a) illustrates, *ControlNet* [48] is prone to generating localized artifacts due to its rigid, pixel-level feature fusion mechanism. The *In-context* [4]



(a) Results of different video-to-video frameworks (b) Ablation study of init-follow mode mechanism.

Fig. 8: Qualitative ablation study of model framework and temporal conditioning.

approach, which incorporates additional video frames, doubles the token sequence length, leading to significantly slower training convergence. More details of this baseline are provided in the Sup. Material. In contrast, our *Multi-Visual-Condition Alignment* design achieves the optimal balance between generative quality and computational efficiency.

Ablation for init-follow mode mechanism. We perform an ablation study to evaluate the impact of temporal priors in follow-mode training. The visual comparisons in Fig. 8(b) demonstrate that without explicit temporal supervision, the model suffers from *temporal discontinuities* at the junction of recurrent segments. Even with a strong foundation model, the instantaneous shift in input triggers undesirable artifacts during the transition. Our init-follow-mode approach, which takes advantage of both temporal conditioning and mask-based temporal encoding, successfully addresses this challenge. This design enables the model to bridge the gap between recurrent segments, leading to superior temporal stability and consistency in extended sequences.

5 Conclusion

We propose EchoStyle, a stable and expandable framework for text-driven stylization of videos with arbitrary lengths. We design a text-driven video-to-video architecture to achieve refined stylistic distillation. Furthermore, we introduce a robust and reusable reverse data generation pipeline to construct a large-scale and high-quality video stylization dataset, *V-Style20k*, which will be open-sourced to the community. To handle long videos, we propose an init-follow-mode mechanism and a sliding-window strategy, scaling video stylization to minute-level duration. Extensive experiments and analysis demonstrate that EchoStyle achieves superior dynamic style consistency as well as fine-grained content preservation for both short and long videos, even comparable to commercial closed-source solutions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62372480).

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Chefer, H., Zada, S., Paiss, R., Ephrat, A., Tov, O., Rubinstein, M., Wolf, L., Dekel, T., Michaeli, T., Mosseri, I.: Still-moving: Customized video generation without customized video data. *ACM Transactions on Graphics (TOG)* **43**(6), 1–11 (2024)
4. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al.: A survey on in-context learning. In: *Proceedings of the 2024 conference on empirical methods in natural language processing*. pp. 1107–1128 (2024)
5. Frenkel, Y., Vinker, Y., Shamir, A., Cohen-Or, D.: Implicit style-content separation using b-lora. In: *European Conference on Computer Vision*. pp. 181–198. Springer (2024)
6. Gao, C., Gu, D., Zhang, F., Yu, Y.: Reconet: Real-time coherent video style transfer network. In: *Asian Conference on Computer Vision*. pp. 637–653. Springer (2018)
7. Gao, J., Sun, Y., Liu, Y., Tang, Y., Zeng, Y., Qi, D., Chen, K., Zhao, C.: Styleshot: A snapshot on any style. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE (2025)
8. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., Li, X., Li, Y., Lin, S., Lin, Z., Liu, J., Liu, S., Nie, X., Qing, Z., Ren, Y., Sun, L., Tian, Z., Wang, R., Wang, S., Wei, G., Wu, G., Wu, J., Xia, R., Xiao, F., Xiao, X., Yan, J., Yang, C., Yang, J., Yang, R., Yang, T., Yang, Y., Ye, Z., Zeng, X., Zeng, Y., Zhang, H., Zhao, Y., Zheng, X., Zhu, P., Zou, J., Zuo, F.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv (Cornell University) (2025). <https://doi.org/https://doi.org/10.48550/arxiv.2506.09113>
9. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: *International Conference on Learning Representations* (2024)
10. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. In: *British Machine Vision Conference* (2017)
11. He, X., Liu, Q., Qian, S., Wang, X., Hu, T., Cao, K., Yan, K., Zhang, J.: Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275 (2024)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: *International Conference on Learning Representations*. vol. 1, p. 3 (2022)
13. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8153–8163 (2024)

14. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
16. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. In: Advances in Neural Information Processing Systems. vol. 38, pp. 70990–71013 (2026)
19. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. Transactions on Machine Learning Research (2024)
20. Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18062–18071 (2022)
21. Li, H., Wang, Y., Huang, T., Huang, H., Wang, H., Chu, X.: Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13684–13694 (2025)
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)
23. Liu, G., Xia, M., Zhang, Y., Chen, H., Xing, J., Wang, Y., Wang, X., Yang, Y., Shan, Y.: Stylecrafter: Enhancing stylized text-to-video generation with style adapter. ACM Transactions on Graphics (2024)
24. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)
25. Liu, Z.S., Wang, L.W., Siu, W.C., Kalogeiton, V.: Name your style: text-guided artistic style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3530–3534 (2023)
26. Mehraban, S., Adeli, V., Rommann, J., Taati, B., Truskovskiy, K.: Pickstyle: Video-to-video style transfer with context-style adapters. arXiv preprint arXiv:2510.07546 (2025)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
28. Qi, T., Fang, S., Wu, Y., Xie, H., Liu, J., Chen, L., He, Q., Zhang, Y.: Deadiff: An efficient stylization diffusion model with disentangled representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8693–8702 (2024)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

30. Shekhar, S., Reimann, M., Hilscher, M., Semmo, A., Döllner, J., Trapp, M.: Interactive control over temporal consistency while stylizing video streams. *Computer Graphics Forum* **42**(4), e14891 (2023)
31. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292* (2024)
32. Song, Q., Lin, M., Zhan, W., Yan, S., Cao, L., Ji, R.: Univst: A unified framework for training-free localized video style transfer. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE (2025)
33. Suresh, A.P., Jain, S., Noinongyao, P., Ganguly, A., Watchareeruetai, U., Samacoits, A.: Fastclipstyler: Optimisation-free text-based image style transfer using style representations. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 7316–7325 (2024)
34. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023)
35. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. *arXiv preprint arXiv:2512.16776* (2025)
36. Tu, S., Xing, Z., Han, X., Cheng, Z.Q., Dai, Q., Luo, C., Wu, Z.: Stableanimator: High-quality identity-preserving human image animation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 21096–21106 (2025)
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025)
38. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733* (2024)
39. Wang, J., Sheng, H., Cai, S., Zhang, W., Yan, C., Feng, Y., Deng, B., Ye, J.: Echoshot: Multi-shot portrait video generation. In: *Advances in Neural Information Processing Systems*. vol. 38, pp. 22058–22090 (2026)
40. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: *ACM SIGGRAPH 2024 Conference Papers*. pp. 1–11 (2024)
41. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025)
42. Xu, J., Mei, Y., Zhang, K., Patel, V.M.: Freevis: Training-free video stylization with inconsistent references. In: *International Conference on Learning Representations* (2026)
43. Yang, Y., Sheng, H., Cai, S., Lin, J., Wang, J., Deng, B., Lu, J., Wang, H., Ye, J.: Echomotion: Unified human video and motion generation via dual-modality diffusion transformer. In: *International Conference on Learning Representations* (2026)
44. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023)
45. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. In: *International Conference on Learning Representations* (2026)
46. Ye, Z., Huang, H., Wang, X., Wan, P., Zhang, D., Luo, W.: Stylemaster: Stylize your video with artistic generation and translation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 2630–2640 (2025)

47. Ye, Z., Liu, Q., Wei, C., Zhang, Y., Wang, X., Wan, P., Gai, K., Luo, W.: Visual-aware cot: Achieving high-fidelity visual consistency in unified models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9116–9126 (June 2026)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
49. Zhang, S., Yang, X., Zi, B., Huang, H., Zhang, C., Li, X.: Telestyle: Content-preserving style transfer in images and videos. arXiv preprint arXiv:2601.20175 (2026)
50. Zhong, Z., Ji, Y., Kong, Z., Liu, Y., Wang, J., Feng, J., Liu, L., Wang, X., Li, Y., She, Y., et al.: Anytalker: Scaling multi-person talking video generation with interactivity refinement. arXiv preprint arXiv:2511.23475 (2025)