

DiT as Real-Time Rerenderer: Streaming Video Stylization with Autoregressive Diffusion Transformer

Hengye Lyu¹, Zisu Li², Yue Hong¹, Yueting Weng³, Jiaxin Shi³,
Hanwang Zhang², and Chen Liang^{1,3*}

¹ The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

hengyelyu@gmail.com, yhong252@connect.hkust-gz.edu.cn,
chenliang2@hkust-gz.edu.cn

² Nanyang Technological University, Singapore, Singapore
{zisu.li, hanwangzhang}@ntu.edu.sg

³ Xmax.AI Ltd., Beijing, China
{yueting, jiaxin}@xmax.ai

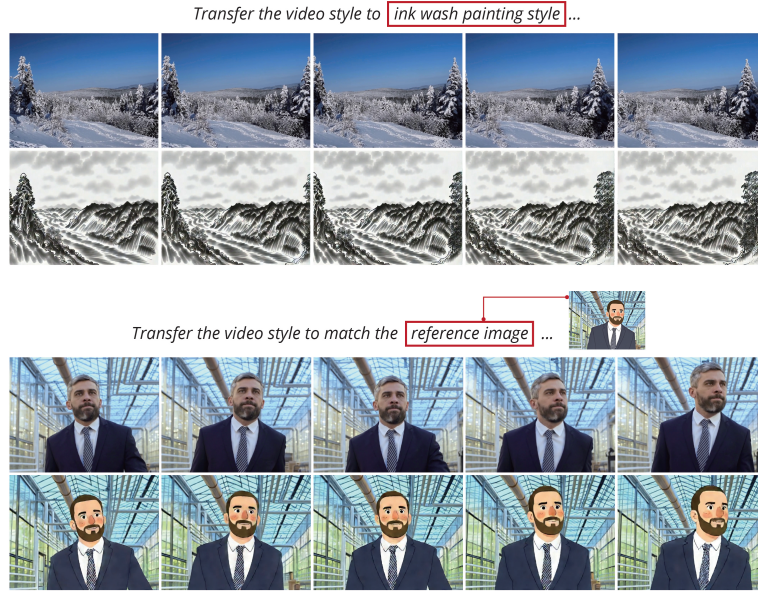


Fig. 1: RTR-DiT is a DiT-based real-time video stylization framework that enables streaming rerendering of video frames, producing stable and consistent results. Top: Text-guided video-to-video (TV2V) stylization results. Bottom: Reference-guided video-to-video (RV2V) stylization results.

* Corresponding author.

Abstract. Recent advances in video generation models have significantly accelerated video generation and related downstream tasks. Among these, video stylization holds important research value in areas such as immersive applications and artistic creation, attracting widespread attention. However, existing diffusion-based video stylization methods struggle to maintain stability and consistency when processing long videos, and their high computational cost and multi-step denoising make them difficult to apply in practical scenarios. In this work, we propose RTR-DiT (DiT as Real-Time Rerenderer), a streaming video stylization framework built upon a Diffusion Transformer. We first fine-tune a bidirectional teacher model on a curated video stylization dataset, supporting both text-guided and reference-guided video stylization tasks, and subsequently distill it into a few-step autoregressive model via post-training with Self Forcing and Distribution Matching Distillation. Furthermore, we propose a reference-preserving KV cache update strategy that not only enables stable and consistent processing of long videos, but also supports real-time switching between text prompts and reference images. Experimental results show that RTR-DiT outperforms existing methods in both text-guided and reference-guided video stylization tasks, in terms of quantitative metrics and visual quality, and demonstrates excellent performance in real-time long video stylization and interactive style-switching applications.

Keywords: Video Stylization · Autoregressive DiT · Real-Time Video Generation

1 Introduction

In recent years, video diffusion models (VDMs) [2, 7, 20] have achieved remarkable progress in high-quality video generation, profoundly advancing research in computer vision and generative artificial intelligence. Leveraging pretrained Text-to-Video (T2V) and Image-to-Video (I2V) diffusion models [31, 36], researchers have further explored a variety of downstream tasks. Among these tasks, video stylization, which aims to transfer a target visual style to an input video while preserving its underlying content and temporal coherence, has attracted increasing attention due to its wide applications in digital content creation, film production, animation, and personalized media generation.

Video stylization based on diffusion models has been extensively researched over a long period. Early methods based on Latent Diffusion Models (LDMs) [24] extended image stylization to video scenes. These methods can generally be divided into two technical approaches: one maps the video into latent space using techniques like DDIM inversion [28] and injects style features, while the other extracts structural and motion information from the scene using frameworks like ControlNet [41], and rerenders based on these priors. Additionally, these methods typically stylize key frames and employ frame propagation strategies to handle long videos [5, 10, 34, 35, 42]. However, LDM-based methods face issues with frame-to-frame inconsistency, and their approach to handling long videos limits

their ability to be applied in real-time scenarios. Recently, Diffusion Transformer (DiT)-based methods have significantly outperformed LDM-based methods in terms of generation quality and consistency. Some works have extended DiT to video stylization tasks by fine-tuning the DiT blocks [37]. Moreover, some works have designed All-In-One video editing frameworks [11, 33] with stylization as a subtask. However, the 3D full attention in DiT blocks is bidirectional in temporal modeling, meaning each frame’s estimation depends on both past and future frames, making it challenging to scale to real-time and long video applications.

Recently, many studies have focused on accelerating diffusion models, making DiT-based models viable for real-time applications. Among these, Distribution Matching Distillation (DMD) [38, 39] distills multi-step sampling diffusion models into single-step or few-step generators, significantly improving generation efficiency while maintaining high output quality. Additionally, some methods have achieved autoregressive DiT models by incorporating causal masks [40], and used techniques such as Diffusion Forcing [3] and Self Forcing [8] to reduce cumulative errors, thereby enhancing the stability and quality of long-video inference. These methods contribute to real-time video generation applications and further extend to Video-to-Video tasks.

In this paper, we propose RTR-DiT (DiT as Real-Time Rerenderer), a real-time video stylization framework built upon an autoregressive DiT. By leveraging the strong generative priors of pretrained video diffusion models, RTR-DiT inherits their powerful capability for high-quality video synthesis, enabling temporally coherent and visually consistent stylized video generation. Moreover, our approach advances video stylization beyond traditional methods that primarily focus on transferring local texture or geometric patterns. Instead, RTR-DiT enables semantically consistent and high-fidelity style transformation, allowing complex visual styles to be applied at the scene and object level while preserving the underlying semantic structure and motion dynamics of the original video. As shown in Fig. 1, RTR-DiT supports both text prompts and reference images as style guidance conditions, processing video frames in a streaming manner to generate high-quality and consistent stylized results.

To implement RTR-DiT, we first constructed a long video stylization dataset containing text-video and reference-video pairs to support the training of our model. Next, we designed a video stylization training and inference framework, which uses text and reference images as conditions to transform the input video into a stylized version. During training, we first fine-tuned a multi-step DiT with temporal bidirectional modeling as the teacher model. Then, in the post-training stage, we distilled it into an autoregressive few-step generator by combining Self Forcing and DMD techniques, enabling real-time stylization for videos. To ensure stable processing of long videos, we designed a reference-preserving KV cache update strategy to improve the consistency of stylization, while also supporting real-time switching between text and reference image conditions, making it more suitable for interactive applications.

In summary, the main contributions of this work are as follows:

- We propose RTR-DiT, the first real-time video stylization framework built on DiT. Conditioned on text prompts and reference images, RTR-DiT performs streaming video-to-video stylization while preserving high visual fidelity and temporal coherence.
- We propose a reference-preserving KV cache update strategy that ensures the autoregressive DiT can stably process video frames during long video stylization, resulting in continuous and condition-consistent stylized outputs.
- We propose an interactive video style-switching method that modifies the video style in real-time by changing the guidance conditions, and extend it to applications in real-world scenarios.

2 Related Work

Video Stylization via Diffusion Models. The success of diffusion models in image stylization has motivated researchers to extend them to the video domain. Some representative works employ U-Net networks to transform video frames in the latent space using text or reference frames [6, 12, 15, 26, 32]. However, these methods often face challenges in maintaining temporal consistency between frames, and thus typically require additional constraints, such as optical flow [4, 5, 10]. Meanwhile, network architectures such as ControlNet [41] enable generation conditioned on scene structures, motion, and other features, motivating several works to explore ControlNet-based video stylization methods [35, 42].

Recently, many video stylization works have been built upon the DiT model. Several works directly train DiT networks to achieve end-to-end style transfer, such as StyleMaster [37] and DreamStyle [16]. Other works have explored video editing frameworks, including stylization [11, 33]. Although these methods achieve significant improvements in video quality, they still face challenges in handling long videos and in real-time scenarios.

Furthermore, existing commercial models [13, 25] only offer video editing APIs with limited duration constraints, failing to meet the demands of long video processing. In contrast, our method can handle longer-duration video editing and supports interactive applications.

Real-Time Video Diffusion Models. Real-time video diffusion models enable low-latency streaming by factorizing synthesis along the temporal dimension. Some methods, such as CogVideo [36], VideoPoet [14], and MAGI [30], adopt autoregressive modeling with spatio-temporal tokens. Although this approach supports streaming generation, purely autoregressive models often struggle to balance visual fidelity and long-range temporal coherence [3, 40]. Moreover, autoregressive video generation introduces a fundamental challenge in long roll-outs due to the training–test mismatch caused by Teacher Forcing, which leads to exposure bias and error accumulation over time. Several forcing paradigms like Diffusion Forcing [3] and Self Forcing [8] have been proposed to address this issue. Diffusion Forcing perturbs temporal context by assigning different noise levels to frames and denoising them jointly, improving robustness to corrupted

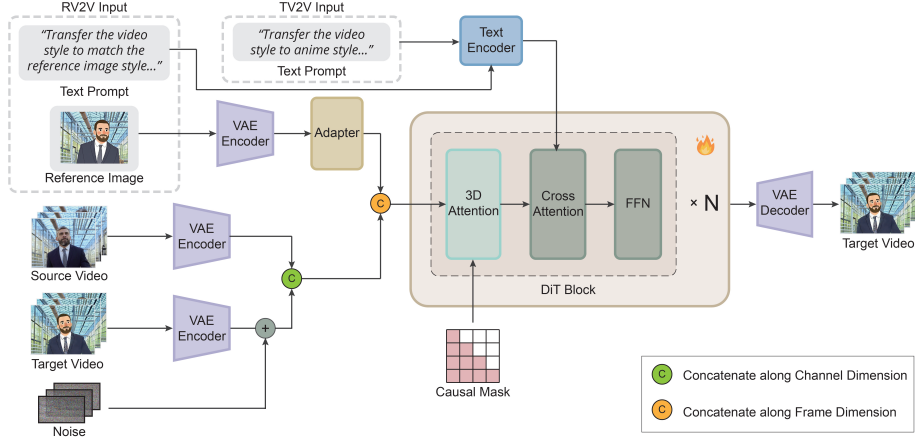


Fig. 2: Overview of RTR-DiT. During training, the original video, the target video, and the reference image are first encoded using a VAE. The noised target video is then concatenated with the source video along the feature dimension, and further concatenated with the reference image along the temporal dimension. When training the model to perform autoregressive prediction, a causal mask is applied within the 3D attention mechanism, enabling the model to perform autoregressive prediction.

histories in streaming settings [40]. Self Forcing further aligns training with inference by conditioning models on self-generated rollouts during training. In parallel, efficiency-oriented techniques such as few-step solvers and distillation-based generators [19, 38, 39] reduce inference latency, but are primarily validated in offline or image-centric scenarios and remain challenging to integrate reliably with long-duration streaming video diffusion. Despite this progress, existing real-time video diffusion models are largely designed for generation rather than video-to-video translation, and struggle to maintain stability and reference consistency under long-horizon streaming inputs. These limitations motivate approaches that explicitly reformulate video stylization as a streaming process with causal temporal modeling and efficient inference.

3 Method

RTR-DiT is an autoregressive DiT model that enables real-time streaming stylization of video frame sequences, guided by prompts or reference images. We refer to this process as "rerendering", which involves editing the appearance and style of the content while preserving the original scene structure and intrinsic attributes. This process includes two modes: Text-guided video-to-video (TV2V) stylization and Reference-guided video-to-video (RV2V) stylization.

3.1 Dataset Creation for Video Stylization

Currently, there is a lack of high-quality, long-duration video stylization datasets that support both text and reference guidance. To address this, we construct a training dataset using existing commercial and open-source large models. We begin by collecting 5,000 videos from the internet, covering a wide range of scenes including human-centric content, animals, and natural scenes. Next, we use Qwen3-VL [1] to generate multiple stylized text prompts and then employ Kling [13] to obtain the corresponding stylized videos. For TV2V mode, we use the prompts Qwen3-VL generated as the text condition. For RV2V mode, we randomly select a frame from the target video as the reference and fix the guiding text to "*Transfer the video style to match the reference image style ...*".

3.2 Preliminary of Base Model

We use the DiT-based Wan [31] video model as the base model. As shown in Fig. 2, during training, we encode the source video V_s and the target video V_t with a variational autoencoder (VAE) [31], obtaining latent representations for N frames, $\mathbf{x}^{1:N} = (\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N)$ and $\mathbf{z}_0^{1:N} = (\mathbf{z}_0^1, \mathbf{z}_0^2, \dots, \mathbf{z}_0^N)$, respectively. If a reference image I_r is provided, we also obtain its VAE latent $\mathbf{c}_{\text{ref}}$. The text prompt is encoded by a text encoder [23] to produce the text conditioning embedding $\mathbf{c}_{\text{text}}$.

Following prior work [8, 40], we first fine-tune a bidirectional DiT as a teacher for our autoregressive DiT. Given a clean latent $\mathbf{z}_0$, we perturb it with noise at a continuous timestep $t \in [0, 1]$ to obtain the noisy latent $\mathbf{z}_t$. We then concatenate $\mathbf{z}_t$ with the source video latent $\mathbf{x}$ along the channel dimension. The source latent $\mathbf{x}$ encodes content and structure cues from the original video, including spatial layout and motion patterns, and provides structural guidance that helps rerendering preserve scene geometry and temporal coherence.

The concatenated latent is then patchified into a sequence of tokens. In RV2V mode, the reference latent $\mathbf{c}_{\text{ref}}$ is processed by a lightweight adapter and prepended to the token sequence. All tokens are subsequently fed into the DiT blocks for modeling.

We fine tune the model using a flow matching [18] objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, c, t} \|\mathbf{v}_\theta(\mathbf{z}_t, \mathbf{x}, c, t) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2, \quad (1)$$

where $\mathbf{z}_0$ denotes the clean latent encoded from data, $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$ is sampled from the standard Gaussian, and

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1, \quad t \sim \mathcal{U}(0, 1). \quad (2)$$

Here $\mathbf{v}_\theta$ denotes the predicted velocity field, and the conditioning c is defined as $c = \{\emptyset, \mathbf{c}_{\text{text}}\}$ for TV2V, or $c = \{\mathbf{c}_{\text{ref}}, \mathbf{c}_{\text{text}}\}$ for RV2V.

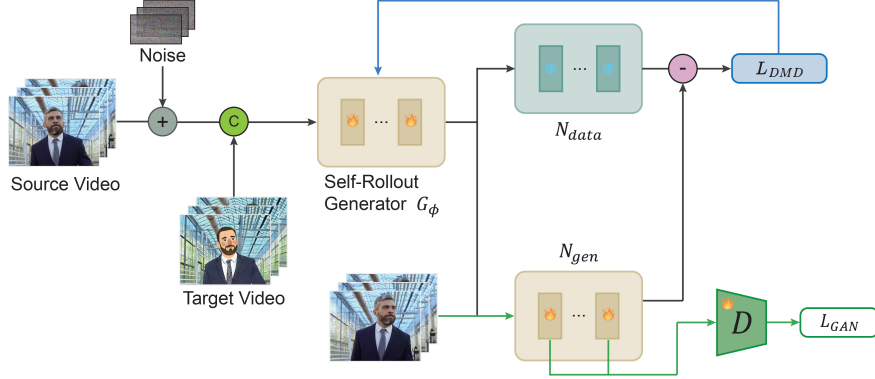


Fig. 3: Post-training pipeline. We distill the bidirectional teacher model into a few-step autoregressive generator using Self Forcing and DMD, and further incorporate adversarial training to improve generation performance.

3.3 Autoregressive DiT for Streaming Stylization

In the teacher model, each block uses full 3D spatiotemporal attention with bidirectional temporal modeling, meaning that the prediction of each frame depends on both past and future frames. This bidirectional dependency is incompatible with streaming video editing. To address this, we convert DiT into an autoregressive form to enable streaming inference and frame-by-frame generation.

Specifically, we impose a temporal causal mask on the 3D attention so that tokens in the noisy latent frame $\mathbf{z}_t^n$ can only attend to tokens from the current frame and preceding frames, *i.e.*, $\{\mathbf{z}_t^1, \mathbf{z}_t^2, \dots, \mathbf{z}_t^n\}$. Let Q , K , and V be the query, key, and value matrices computed from the token sequence of $\mathbf{z}_t$, the masked attention is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right)V, \quad (3)$$

$$\text{where } M^{ij} = \begin{cases} 0, & \lfloor j/k \rfloor \leq \lfloor i/k \rfloor, \\ -\infty, & \text{otherwise.} \end{cases}$$

Here M is the causal mask, d is the attention head dimension, and k denotes the number of spatial tokens per latent frame. This construction ensures that each token attends only to tokens from the current and past frames.

Moreover, to improve generation efficiency and enable few-step inference, we perform denoising along a fixed timestep sequence $\{s_1, s_2, \dots, s_K\}$, where $s_i \in [0, 1]$, instead of iterating densely over the entire time horizon. Under this setting, we obtain a few-step autoregressive generator G_ϕ , which is further optimized through post-training.

Self-Rollout. We adopt Self Forcing training, where the model conditions each frame generation step on its own previously generated frames rather than ground-truth history, thereby reducing the train–test gap in autoregressive inference. Formally, the conditional generation distribution of G_ϕ under Self Forcing is

$$p_\phi(\hat{\mathbf{z}}_0^{1:N} \mid \mathbf{x}^{\leq N}, c) = \prod_{n=1}^N p_\phi(\hat{\mathbf{z}}_0^n \mid \hat{\mathbf{z}}_0^{\leq n}, \mathbf{x}^{\leq n}, c). \quad (4)$$

Distribution Matching Distillation. We adopt Distribution Matching Distillation (DMD) [38, 39] to distill the bidirectional, multi-step teacher model into G_ϕ . As shown in Fig. 3, we initialize two score networks with the teacher’s weights: a frozen network N_{data} for estimating the score of the real data distribution, and a trainable network N_{gen} for estimating the score of the generator-induced distribution. DMD aims to minimize the reverse Kullback-Leibler (KL) divergence between the real and generated distributions. Its gradient can be expressed in terms of two score functions as follows:

$$\begin{aligned} \nabla_\phi \mathcal{L}_{\text{DMD}} = \mathbb{E}_{\mathbf{z}_t, \mathbf{x}, c, t} \Big[& - (S_{\text{data}}(F(G_\phi(\mathbf{z}_t, \mathbf{x}, c), t), t) \\ & - S_{\text{gen}}(F(G_\phi(\mathbf{z}_t, \mathbf{x}, c), t), t)) \frac{dG}{d\phi} \Big] \end{aligned} \quad (5)$$

where F is forward diffusion process, and S_{data} and S_{gen} are score functions [29] estimated by N_{data} and N_{gen} , respectively. After each update of G_ϕ , we update N_{gen} five times to match the score of the current generator distribution [40].

Adversarial Post-Training. Moreover, we adopt adversarial training [17, 39] to further improve the generation quality of G_ϕ . We attach an additional classification branch to N_{gen} as a discriminator D , and optimize the G_ϕ and D using a standard GAN objective:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{\mathbf{z}_0, \mathbf{x}, c} [\log D(\mathbf{z}_0, \mathbf{x}, c)] + \mathbb{E}_{\mathbf{z}_t, \mathbf{x}, c, t} [-\log D(F(G_\phi(\mathbf{z}_t, \mathbf{x}, c), t), \mathbf{x}, c)] \quad (6)$$

3.4 Inference for Consistent Long-Video Generation

To enable continuous editing of arbitrarily long videos, we follow the approach of Huang *et al.* [8] and adopt a rolling KV cache mechanism. Specifically, we initialize fixed-length KV caches for both 3D attention and cross attention. During inference, newly generated frames are appended to the KV cache of the 3D attention, and the cache is updated in a rolling manner. Once the cache reaches its maximum capacity, the oldest frames are removed to make room for new ones.

In RV2V mode, a simple strategy is to prepend the reference image before the first frame and update the KV cache in a rolling manner. However, once the

cache reaches its maximum capacity, the reference image is removed, preventing subsequent frames from attending to it and leading to noticeable degradation.

To address this issue, we propose a reference-preserving KV cache update strategy. After the cache is fully populated, the reference image is permanently anchored at the front of the cache, while the remaining positions continue to roll forward as new frames are generated. This design ensures that all subsequent frames can consistently attend to the reference image, thereby enabling stable long-video generation.

Furthermore, we modify the KV cache update mechanism to enable real-time switching of video styles by changing the text prompt or reference image. When a new text prompt or reference image is provided to the ongoing stylization model, the KV cache of the cross attention is reinitialized. Meanwhile, the 3D attention KV cache retains the most recent frame to ensure a smooth transition during the style change. If the new input is a reference image, it is anchored at the front of the KV cache so that subsequent frames can consistently attend to the reference style. This strategy enables more flexible deployment in real-time interactive scenarios.

4 Experiments

4.1 Experimental Settings

Training Details. RTR-DiT is trained in three stages. First, we fine-tune a teacher model with bidirectional temporal modeling on our constructed dataset for 10,000 steps. Next, we initialize a few-step student model using a fixed timestep sequence $\{s_1, s_2, \dots, s_K\}$ and train it for 10,000 steps to obtain a strong initialization. Finally, we apply post-training distillation to obtain the final model, trained for 5,000 steps. In our experiments, we set $K = 2$, which provides the best trade-off between generation quality and inference speed. During training, we jointly train on a mixture of TV2V and RV2V samples.

Evaluation dataset. We collect 50 videos from the Pexels website [21] to build an evaluation benchmark. In TV2V mode, we prepare 10 diverse stylization prompts (e.g., ink wash painting and Minecraft styles) and evaluate the task using random text-video pairings from the benchmark. In the RV2V mode, to simplify the setup and enable standardized comparison with other reference-guided methods, we extract the first frame of each video and apply random style transformations using Kling [13] as the reference. More RV2V stylization results can be found in the supplementary materials. All videos are resized to 5-second clips with a resolution of 832×640 and a frame rate of 24 fps. Except for the commercial models, all tests are conducted on a single H20 GPU.

4.2 Comparison with Text-Guided Video Stylization Methods

We compare our approach with three diffusion-based stylization methods: Rerender-A-Video [34], FRESCO [35], and TokenWarping [42]. Unlike our instructional

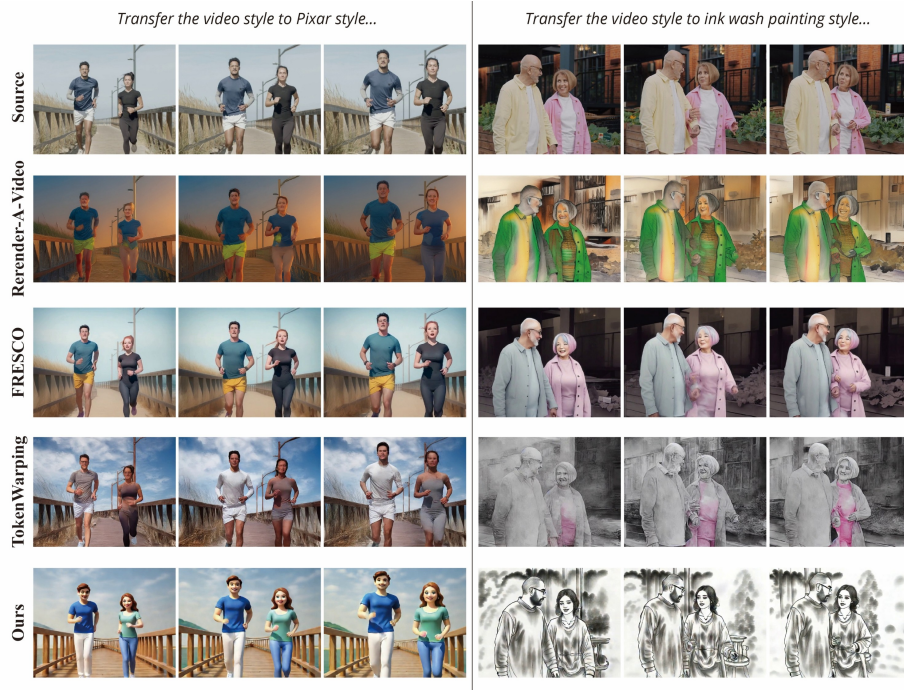


Fig. 4: Visual comparison with TV2V stylization methods. Our method generates stylized videos that are more consistent with the text prompt and achieves superior visual quality compared to other methods.

prompts, these three methods use descriptive prompts. We use Qwen3-VL [1] to generate descriptive prompts for the stylized scene content. For quantitative evaluation, we use the following metrics: CLIP-T [22] to measure the text-video similarity via CLIP, CLIP-F to evaluate temporal consistency through intra-frame similarity, and AestheticQuality(AQ) and ImagingQuality(IQ) from VBench [9] to assess video quality. Additionally, we also measure the generation time for each video.

Tab. 1 presents a comparison of our method with three other text-guided video stylization approaches. The results show that RTR-DiT achieves the best performance on CLIP-T, indicating that our method better follows the text guidance. It also performs the best on CLIP-F, demonstrating superior temporal consistency compared to the other methods. Additionally, our method achieves the highest AestheticQuality and second-best ImagingQuality, indicating that the generated video quality is notably superior. In terms of generation time, our method significantly outperforms the other three approaches, highlighting the efficiency advantage of autoregressive models in streaming stylization.

Fig. 4 shows a comparison of the visual effects of the four methods. When converting to the ink wash painting style, the other methods incorrectly alter the

video’s color tones while maintaining the scene’s structure, whereas our method accurately transforms the video into the ink wash style. When converting to Pixar style, our method generates character designs that are closer to Pixar animation characters compared to the other methods. These results demonstrate the significant advantages of our method in text-guided video stylization.

4.3 Comparison with Reference-Guided Video Stylization Methods

We selected two DiT-based methods: VACE [11] and StyleMaster [37], as well as the commercial model Gen-4 Aleph [25], for comparison. In the evaluation metrics, we use CSD-Score [27] to replace CLIP-T, which measures the similarity between reference images and video frames.

Tab. 1 presents a comparison of our method with three other reference-guided video stylization approaches. Our method achieves the best results in Aesthetic-Quality, and second-best in CLIP-F, CSD-Score and ImagingQuality, just behind the commercial model Gen-4 Aleph. This indicates that RTR-DiT outperforms existing open-source methods in reference-guided character stylization and is competitive with commercial models.

Fig. 5 presents a qualitative comparison of the four methods. VACE, when generating stylized videos guided by reference images, only adjusts the image brightness and fails to capture the style features effectively. The stylized videos generated by StyleMaster show significant visual style differences compared to the reference image. Both Gen-4 Aleph and our method generate high-quality stylized results, but our results are closer to the reference image and more stable.

In addition, VACE and StyleMaster struggle with long inference times, which limits their applicability in real-time scenarios. Although the commercial model Gen-4 Aleph achieves shorter inference times than the former two methods, users still have to endure long queuing times in practice. In contrast, our method maintains fast generation while still producing high-quality results.

Table 1: Quantitative comparison results. Top: TV2V methods. Bottom: RV2V methods. Our method demonstrates strong performance across all metrics in both TV2V and RV2V settings.

Methods	CLIP-T/CSD-Score	CLIP-F	AQ	IQ	Time(min)
Rerender-A-Video	0.2272	0.9883	0.5730	0.7012	4.5
FRESCO	0.1970	0.9907	0.6060	0.6834	7.3
TokenWarping	0.2122	0.9857	0.6093	0.7324	3.2
Ours	0.2585	0.9914	0.6302	0.7279	0.12
VACE	0.4046	0.9905	0.6012	0.7165	27.6
StyleMaster	0.4275	0.9892	0.6536	0.7247	28.5
Gen-4 Aleph	0.8312	0.9925	0.6336	0.7378	3.5
Ours	0.7958	0.9909	0.6714	0.7321	0.12

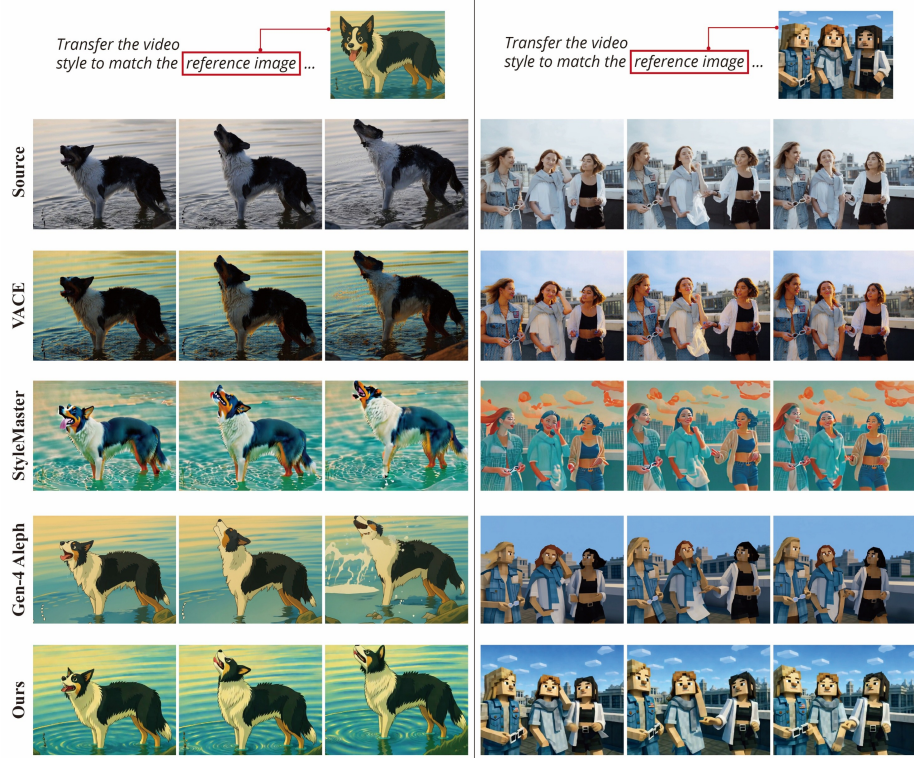


Fig. 5: Visual comparison with RV2V stylization methods. Our method achieves performance comparable to commercial models under real-time inference conditions.

4.4 Ablation Study

We first evaluate the effectiveness of post-training by comparing the performance of models obtained at different training stages. The corresponding quantitative metrics and visual comparisons are shown in Tab. 2 and Fig. 6. The bidirectional teacher model achieves strong performance across multiple metrics and produces high-quality visual results. In contrast, the causal model without post-training shows a noticeable drop in quantitative metrics, and the quality of generated frames gradually degrades as inference proceeds. After post-training, our model achieves results that are close to or even surpass the teacher model on several metrics, while producing visual results comparable to those of the teacher model.

We further validate the effectiveness of the proposed reference-preserving (RP) KV cache update strategy through qualitative evaluation. As shown in Fig. 6b, without fixing the reference tokens at the front of the cache, the generated frames gradually degrade during inference and drift away from the style of the reference image. In contrast, our method enables the model to generate videos with stable and consistent stylization.

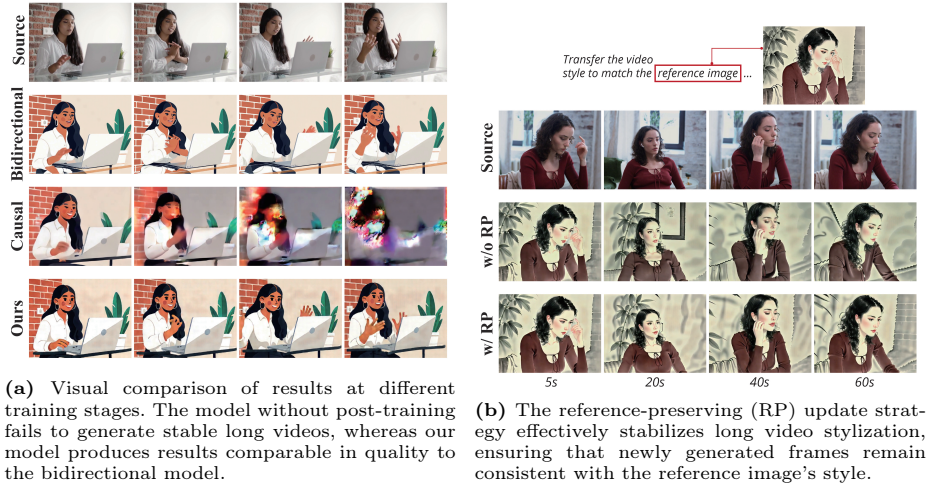


Fig. 6: Qualitative comparison results of ablation studies.

Table 2: Quantitative comparison across different training stages. Top: TV2V mode. Bottom: RV2V mode. After post-training, our method not only achieves fast inference but also shows significant improvements across all metrics.

Methods	CLIP-T/CSD-Score	CLIP-F	AQ	IQ	Time(min)
Bidirectional	0.2463	0.9931	0.6514	0.7406	1.43
Causal	0.2284	0.9948	0.5446	0.6380	0.12
Ours	0.2585	0.9914	0.6302	0.7279	0.12
Bidirectional	0.8050	0.9930	0.6583	0.7144	1.43
Causal	0.6002	0.9956	0.5897	0.6021	0.12
Ours	0.7958	0.9909	0.6714	0.7321	0.12

4.5 User study

We conduct a user study on both TV2V and RV2V with 12 participants and 16 randomized comparison groups, 8 per setting. Participants rank four anonymized methods under three criteria: condition/style alignment, visual quality, and overall preference. Tab. 3 reports the average rank, where lower is better. RTR-DiT ranks first on all TV2V criteria, and achieves the best alignment and overall preference in RV2V, with visual quality comparable to the strongest baseline.

4.6 Long Video Stylization and Applications

We collect several long videos from real-world scenes to demonstrate the ability of RTR-DiT in stable long-video stylization. As shown in Fig. 7, our model is capable of maintaining stylization over long videos of approximately 1-minute duration, with the generated video consistently matching the style of the reference

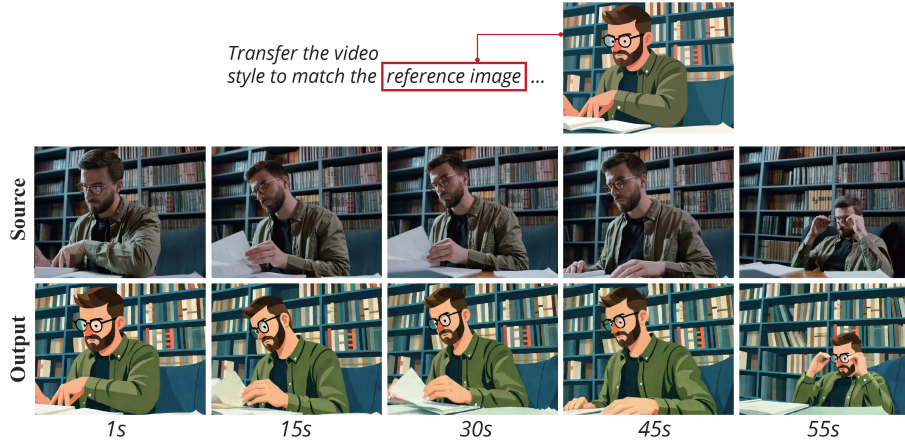


Fig. 7: Long video stylization example. RTR-DiT maintains stable stylization over extended durations under streaming input.

image. This is attributed to the enhancement of autoregressive DiT capabilities through post-training and the improvement in inference performance brought by the reference-preserving update strategy. More long-video stylization examples can be found in the supplementary materials.

We also demonstrate the ability to interactively switch the style of a video in real time by modifying the text or reference. As shown in Fig. 8, the video is initially guided by an animated reference image, followed by a switch to a text-based prompt for an ink painting style, and finally, another text-based prompt is used to switch the style to Impressionism. RTR-DiT enables style changes under different guidance conditions, and the generated results remain faithful to the guidance conditions while maintaining high quality and consistency. This demonstrates the significant research potential of our method in real-time immersive applications such as VR/AR. Based on RTR-DiT, we have developed a real-time

Table 3: User study results on TV2V (top) and RV2V (bottom).

Method	Align. ↓	Quality ↓	Overall ↓
Rerender-A-Video	2.78	3.05	2.91
FRESCO	3.25	2.48	2.74
TokenWarping	2.83	3.15	3.10
RTR-DiT	1.14	1.32	1.25
VACE	3.88	3.20	3.48
StyleMaster	3.07	3.58	3.45
Gen-4 Aleph	1.76	1.60	1.64
RTR-DiT	1.29	1.61	1.44

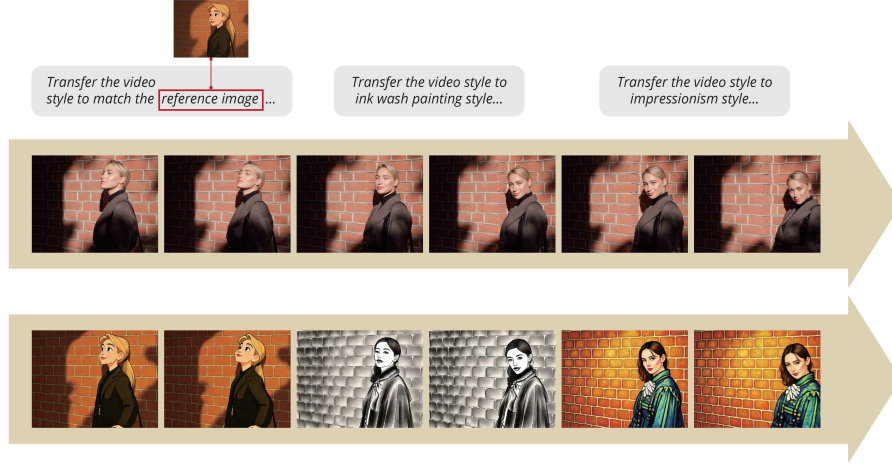


Fig. 8: Text/reference-guided switching example. Our model can switch the guidance conditions in real-time.

stylization application that uses a smartphone as the front end, streaming the captured video to a network deployed on a server and outputting the stylized results in real time. Further details and additional style-switching examples can be found in the supplementary materials.

5 Conclusion

In this paper, we propose RTR-DiT, a real-time video stylization framework based on DiT that enables streaming, high-quality, and spatiotemporally consistent video stylization through both text and reference image guidance. To achieve this, we construct a high-quality long video stylization dataset containing text-video and reference-video pairs, and train an autoregressive generator through a staged approach. We first fine-tune a bidirectional DiT as the teacher model, then design a self-rollout causal DiT as the student model, and use the DMD algorithm for distillation training. Additionally, we introduce a reference-preserving KV cache update strategy to enable stable and consistent processing of long videos, while supporting real-time switching between prompts and reference images. Experimental results show that our method outperforms existing open-source approaches in both TV2V and RV2V modes in terms of quantitative metrics and visual quality, and is competitive with commercial models.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant No. 62502411.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
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. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems* **37**, 24081–24125 (2024)
4. Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: Flatten: optical flow-guided attention for consistent text-to-video editing. arXiv preprint arXiv:2310.05922 (2023)
5. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
6. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
7. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *Advances in neural information processing systems* **35**, 8633–8646 (2022)
8. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
9. 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)
10. Jeong, H., Ye, J.C.: Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. arXiv preprint arXiv:2310.01107 (2023)
11. 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)
12. Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6507–6516 (2024)
13. Kling AI: Kling AI: Next-generation ai creative studio. <https://kling.ai/a> (2026), accessed: 2026-06-25
14. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023)
15. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468 (2024)
16. Li, M., Chen, J., Zhao, S., Feng, W., Tu, P., He, Q.: DreamStyle: A unified framework for video stylization. arXiv preprint arXiv:2601.02785 (2026)
17. Lin, S., Xia, X., Ren, Y., Yang, C., Xiao, X., Jiang, L.: Diffusion adversarial post-training for one-step video generation. arXiv preprint arXiv:2501.08316 (2025)

18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
19. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
20. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
21. Pexels: Pexels. <https://www.pexels.com/> (2025), accessed: 2025-12-14
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
23. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)
24. 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)
25. Runway: Introducing runway gen-4 (2026), <https://runwayml.com/research/introducing-runway-gen-4>, accessed: 2026-01-20
26. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
27. 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)
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
29. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
30. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
31. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
32. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: 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. pp. 7623–7633 (2023)
33. Wu, Y., Chen, L., Li, R., Wang, S., Xie, C., Zhang, L.: Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16692–16701 (2025)

34. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)
35. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Fresco: Spatial-temporal correspondence for zero-shot video translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8703–8712 (2024)
36. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
37. 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)
38. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems* **37**, 47455–47487 (2024)
39. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
40. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
41. 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)
42. Zhu, H., Xu, Y., Yu, J., He, S.: Zero-shot video translation via token warping. *IEEE Transactions on Visualization and Computer Graphics* (2025)