STREAMING VIDEO EDITING WITH EASY ADAPTATION

Anonymous authors
Paper under double-blind review

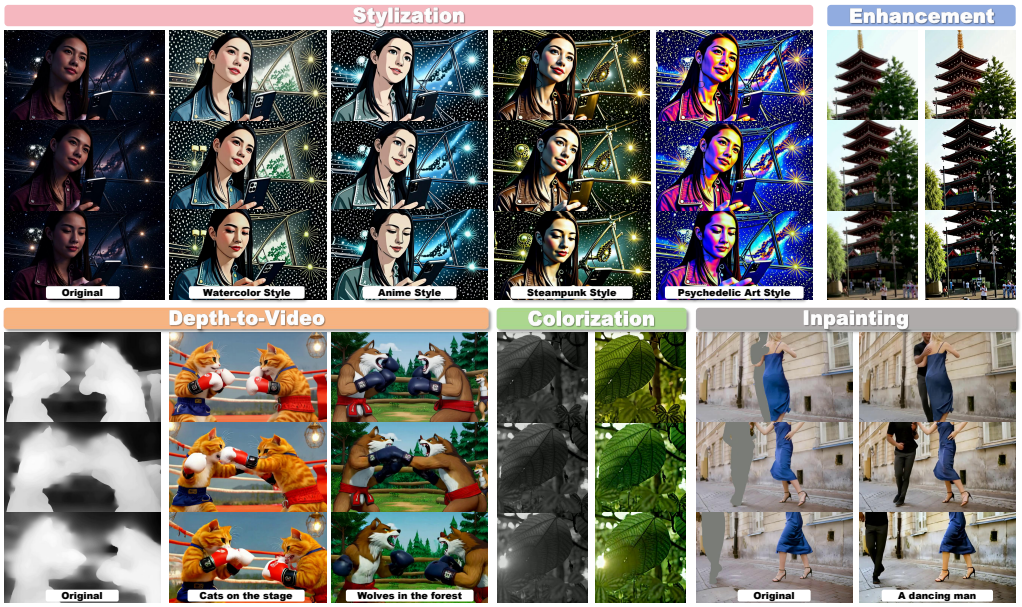


Figure 1: Visualization of the proposed SVEET. Our method can perform high-quality streaming video editing using only bidirectional model training. Here we demonstrate the results generated by SVEET on five video editing tasks, including stylization, enhancement, depth-to-video, colorization and inpainting.

ABSTRACT

In this paper, we propose SVEET, a framework that requires merely training on a pretrained bidirectional video diffusion model but supports high-quality streaming video editing in an auto-regressive fashion. To tackle this problem, we first systematically revisit existing video-to-video diffusion approaches and identify two key principles for such streaming adaptation: backbone feature disentanglement and conditional frame independence. Building on these insights, we develop a novel paradigm for controllable video generation. At its core, an auxiliary model branch encodes source video inputs with temporally independent self-attention, and the intermediate features are injected into the corresponding backbone blocks for streaming-compatible control. Moreover, to bridge the discrepancy between the feature spaces of bidirectional and streaming models, we propose a decoupled training scheme that explicitly enforces the orthogonality between the optimization directions of video controllability and model causality. Such disentanglement ensures compatibility between the two objectives at inference and facilitates smooth zero-shot knowledge transfer across heterogeneous backbone architectures. Extensive experiments demonstrate that SVEET achieves superior editing quality while maintaining real-time performance, attaining 15 FPS on a single H100 GPU without any auxiliary acceleration techniques.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Driven by the growing demand for real-time and interactive applications, video diffusion models Blattmann et al. (2023); Ho et al. (2022); Kong et al. (2024); Liu et al. (2024); Wan et al. (2025) are increasingly moving from offline generation toward streaming settings Yang et al. (2025); Kodaira et al. (2025); Feng et al. (2025), where frames are generated causally to support low-latency applications such as interactive content creation and real-time avatar synthesis Mahmoud & Abozariba (2025); Lu et al. (2025b). Recent approaches Yin et al. (2025b); Huang et al. (2025); Zhu et al. (2026) typically adapt pretrained bidirectional Diffusion Transformers (DiTs) into causal autoregressive models, often together with distillation for efficient sequential generation.

Despite this progress, DiT-based streaming video editing remains largely unexplored. Existing video editing methods Kong et al. (2024); Liu et al. (2024); Wan et al. (2025) rely on bidirectional full-sequence processing, which fundamentally violates streaming constraints so that making them incompatible with causal streaming applications such as live style transfer and online inpainting. To bridge this gap, a straightforward solution is to train a streaming editing model from scratch. However, this is prohibitively expensive, as pretraining a competitive video diffusion backbone requires massive computational resources. According to recent works Yin et al. (2025a); Huang et al. (2025); Zhu et al. (2026), such streaming distill can take 128 H100 GPU days and involve synthesizing thousands of ODE pairs. Moreover, incorporating control signals further increases training cost.

Motivated by these inconveniences, in this paper, we are curious about one interesting question: *Is it possible to achieve high-quality streaming video editing by training solely on a pretrained bidirectional model and directly transfer the learned editing capability to a streaming setting without any retraining?* To investigate this problem, we first systematically revisit existing video-to-video diffusion architectures and identify two fundamental principles for such streaming-compatible control: (1) *Backbone feature disentanglement*, where the control mechanism must remain decoupled from the base model to preserve the encapsulated pretrained knowledge; and (2) *Conditional frame independence*, where source video encoding must be performed in a per-frame basis without temporal dependency to ensure causal compatibility during streaming inference.

Guided by these insights, we propose *SVEET*, a framework which enables streaming video editing with easy adaption and answer the above question positively. At its core, SVEET introduces an auxiliary branch that encodes source video inputs using temporally independent self-attention and injects the resulting features into corresponding backbone blocks for streaming-compatible control. We verify that such an architecture adheres to the above guidelines and achieves plausible results in the zero-shot transfer from bidirectional to streaming settings.

Nevertheless, a fundamental issue remains: the feature spaces of bidirectional and autoregressive models are not fully aligned, which hinders zero-shot transfer. Regarding this, we further propose a decoupled training scheme that explicitly enforces the orthogonality between optimization directions for video controllability and model causality. Specifically, we first use singular value decomposition (SVD) Stewart (1993); Wall et al. (2003) to extract the dominant update directions capturing the discrepancy between autoregressive and bidirectional models and then constrain the control branch to optimize in the orthogonal subspace. We theoretically validate that such disentanglement facilitates the compatibility between the two objectives and enables effective zero-shot transfer across heterogeneous backbones.

Extensive experiments across multiple editing tasks show that SVEET achieves strong editing quality while running at 15 FPS on a single H100 GPU without auxiliary acceleration. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to study bidirectional-to-streaming video editing transfer without retraining or distilling the streaming backbone.
- We identify two principles for streaming-compatible transfer—backbone feature disentanglement and conditional frame independence—and develop SVEET, a temporally independent control architecture that directly transfers from a bidirectional to a streaming backbone.
- We introduce Orthogonal Decoupled Training (ODT) to mitigate the bidirectional-to-causal feature mismatch by decoupling controllability learning from causal dynamics. Extensive experiments demonstrate effective zero-shot transfer across diverse video editing tasks while enabling real-time streaming inference.

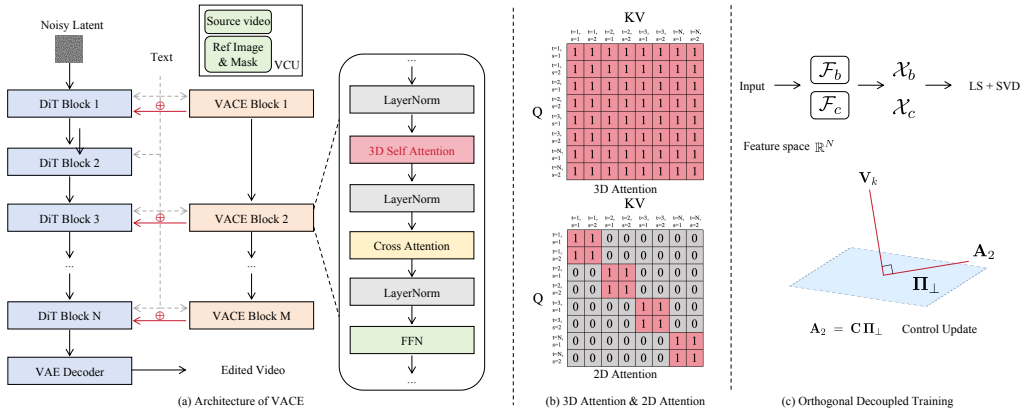


Figure 2: Our proposed SVEET. Our method builds upon VACE (a) and introduces a temporally independent 2D attention design in the control branch (b). To enable effective transfer from bidirectional to streaming backbones, we further incorporate an orthogonal decoupled training scheme (c) to bridge the feature discrepancy between the two settings.

2 Related Work

2.1 Video Editing with Diffusion Models

Recent video diffusion models Blattmann et al. (2023); Ho et al. (2022); Kong et al. (2024); Liu et al. (2024); Wan et al. (2025) support a broad range of editing tasks, including appearance editing Ding et al. (2023); Bai et al. (2025a), structure-guided generation Xing et al. (2024), and object-level manipulation Ma et al. (2025). Early methods Lei et al. (2025); Wu et al. (2023); Xu et al. (2023); Wang et al. (2024) extend image diffusion models Ho et al. (2020); Song et al. (2020); Rombach et al. (2022) with temporal modules or attention propagation, while recent works Jiang et al. (2025); Cheng et al. (2023); Bai et al. (2025b) increasingly adopt unified conditional architectures for flexible multi-task editing. However, most existing approaches rely on bidirectional full-sequence processing, limiting their applicability to real-time streaming scenarios.

2.2 Real-time Streaming Video Generation

Recent real-time video generation methods mainly pursue two directions: causal autoregressive generation Sun et al. (2019); Yan et al. (2021); Singer et al. (2022); Villegas et al. (2022); Henschel et al. (2025) and inference acceleration through consistency or distillation Song et al. (2023); Luo et al. (2023); Zheng et al. (2025); Yin et al. (2024); Salimans & Ho (2022); Lu et al. (2025a). Recent approaches Huang et al. (2025); Zhu et al. (2026); Zhao et al. (2026) further combine autoregressive modeling with diffusion distillation to reduce the gap between training and inference and enable efficient streaming generation. However, these methods primarily target video generation rather than video editing and often require costly adaptation or distillation. In contrast, we study transferring editing capability learned on a bidirectional model directly to a pretrained streaming backbone without retraining it.

3 Methods

3.1 Preliminaries

Conditional Video Diffusion Architecture. Introducing a plug-and-play control module has become a common practice in the field of controllable generation. For instance, in video domain, Wan-VACE Jiang et al. (2025) extends Wan-T2V Wan et al. (2025) with a separate control branch for unified video editing. As shown in Fig. 2(a), the branch encodes multimodal conditions through a Video Condition Unit (VCU) and injects the resulting features additively into intermediate blocks of a frozen DiT backbone, which enables unified controllable video generation and editing without compromising the pretrained generative prior.

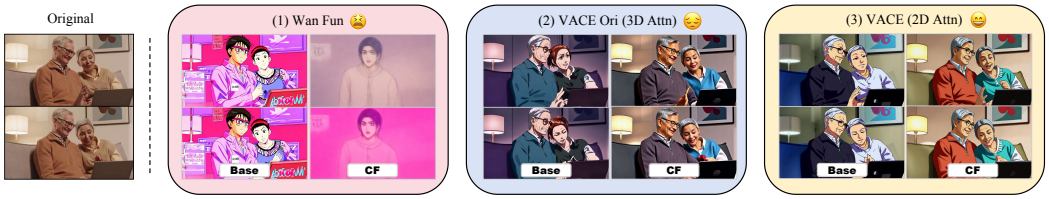


Figure 3: Performance of three representative structures during bidirectional-to-stream transfer.

Autoregressive Video Diffusion Pipeline. Unlike bidirectional models that jointly denoise all frames, autoregressive (AR) video diffusion models generate latent chunks $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K]$ causally:

$$p_\theta(\mathbf{z} \mid \mathbf{c}) = \prod_{k=1}^{K} p_\theta(\mathbf{z}_k \mid \mathbf{z}_{<k}, \mathbf{c}), \tag{1}$$

where $\mathbf{c}$ denotes the text condition. During inference, each chunk is denoised using the accumulated key/value cache of preceding chunks:

$$\mathbf{z}_k^{(t-1)} \;=\; \mathcal{F}_\theta\Big(\mathbf{z}_k^{(t)},\, \mathbf{KV}_{<k},\, \mathbf{c},\, t\Big), \qquad t = T, T-1, \ldots, 1, \tag{2}$$

which is then updated for subsequent generation.

Bridging Bidirectional and Streaming Video Diffusion. To investigate how video editing capability transfers from bidirectional diffusion models to streaming autoregressive backbones, we conduct comparative experiments on three representative architectures: (1) Wan-Fun, which fine-tunes backbone parameters during training; (2) VACE with full spatiotemporal attention; and (3) VACE with temporally independent 2D attention by transferring their learned editing capability from a bidirectional to a streaming backbone.

As shown in Fig. 3, VACE-Fun suffers severe degradation after transfer, almost completely failing to transfer, full spatiotemporal VACE retains only partial editing capability with unstable temporal behavior, whereas the 2D-attention variant transfers substantially more reliably.

These observations suggest two key principles for bridging bidirectional and streaming video diffusion: (P1) Backbone feature disentanglement: control learning should remain decoupled from the bidirectional backbone; and (P2) Conditional frame independence: source-video conditioning should avoid cross-frame dependencies that conflict with causal chunk-wise inference. These principles motivate the design of SVEET.

3.2 SVEET

Problem Statement. We consider two pretrained video diffusion backbones: a bidirectional model $\mathcal{F}_b$, which jointly denoises all video frames, and a causal streaming model $\mathcal{F}_c$, which generates videos sequentially in a chunk-wise autoregressive manner as introduced in Sec. 3.1. Our goal is to learn video editing capability on the frozen $\mathcal{F}_b$ and directly transfer it to the frozen $\mathcal{F}_c$ without retraining or fine-tuning the streaming backbone.

This transfer presents two challenges: the conditioning pathway must remain compatible with causal inference, while the feature spaces of the two backbones are not perfectly aligned. Accordingly, SVEET consists of two complementary components: (i) a temporally independent control branch for causal-compatible conditioning, and (ii) Orthogonal Decoupled Training (ODT) for mitigating the bidirectional-to-causal feature discrepancy.

Temporally Independent Control. We build the control pathway upon a VACE-like conditional video diffusion backbone, which already satisfies backbone feature disentanglement. Motivated by the conditional frame independence principle identified in our analysis, we further modify the control branch by replacing the original full spatiotemporal self-attention with temporally independent 2D spatial attention to ensure frame-wise independent encoding, as shown in Fig. 2(b). Concretely, denoting each token by its frame index t and spatial index s, the attention mask is defined as

$$M_{(t,s),(t',s')} \;=\; \begin{cases} 1, & \text{if } t = t'; \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

such that each query attends only to tokens from the same source frame, reducing the attention to a per-frame 2D spatial self-attention. Consequently, the conditioning representation of frame t depends only on that frame, while temporal coherence of the generated video remains modeled by the causal streaming backbone through its autoregressive history. As a direct consequence, the control branch requires no key/value cache during streaming inference, avoiding introducing any additional cache-growth overhead on top of the AR backbone in Eq. 2.

Orthogonal Decoupled Training. Although the control branch is trained on a bidirectional diffusion backbone $\mathcal{F}_b$, it is deployed on a causal streaming backbone $\mathcal{F}_c$ with different parameters and feature distributions. This mismatch leads to degraded transfer of learned control signals when directly applied in streaming settings. We address this issue by explicitly decoupling controllability learning from the bidirectional-to-causal feature discrepancy.

We first probe the feature gap in a data-driven way. Given a small calibration set of video–prompt pairs, we feed identical inputs to $\mathcal{F}_b$ and $\mathcal{F}_c$ and record the per-block hidden states $\mathbf{X}_b^{(\ell)}, \mathbf{X}_c^{(\ell)} \in \mathbb{R}^{n\times N}$, where N is the total number of cached tokens at block ℓ. A closed-form linear map between the two feature streams is obtained via ridge regression, which corresponds to a regularized least-squares estimation:

$$\mathbf{W}_1^{(\ell)} = \arg\min_{\mathbf{W}} \left\| \mathbf{W}\mathbf{X}_b^{(\ell)} - \mathbf{X}_c^{(\ell)} \right\|_F^2 + \lambda \|\mathbf{W}\|_F^2, \tag{4}$$

where λ is the ridge regularization coefficient. Taking the bidirectional feature itself as the reference, i.e., $\mathbf{W}_b = \mathbf{I}$, the residual transformation:

$$\mathbf{A}_1^{(\ell)} \;=\; \mathbf{W}_1^{(\ell)} - \mathbf{I} \;\in\; \mathbb{R}^{n\times n} \tag{5}$$

captures, in feature space, the *update direction* that turns a bidirectional representation into its causal counterpart at block ℓ. To localize the directions along which $\mathbf{A}_1^{(\ell)}$ acts most strongly, we apply a thin singular value decomposition

$$\mathbf{A}_1^{(\ell)} \;=\; \mathbf{U}^{(\ell)}\, \mathbf{\Sigma}^{(\ell)}\, \mathbf{V}^{(\ell)\top}, \tag{6}$$

and retain the top-k right singular vectors $\mathbf{V}_k^{(\ell)} \in \mathbb{R}^{n\times k}$ associated with the k largest singular values. The columns of $\mathbf{V}_k^{(\ell)}$ form an orthonormal basis of the *discrepancy subspace*, capturing the dominant directions of the bidirectional-to-causal feature shift. We empirically verify that the discrepancy energy is concentrated in a compact set of singular directions across DiT layers in Appendix A. The orthogonal projector onto its complement is

$$\mathbf{\Pi}_\perp^{(\ell)} \;=\; \mathbf{I} - \mathbf{V}_k^{(\ell)}\mathbf{V}_k^{(\ell)\top}, \tag{7}$$

which satisfies $\mathbf{\Pi}_\perp^{(\ell)}\mathbf{V}_k^{(\ell)} = \mathbf{0}$ by construction.

At each block ℓ, the control branch introduces a learnable update $\mathbf{A}_2^{(\ell)}$ to the backbone feature transformation. To prevent this update from relying on the dominant bidirectional-to-causal discrepancy directions, we constrain it as

$$\mathbf{A}_2^{(\ell)} = \mathbf{C}^{(\ell)}\mathbf{\Pi}_\perp^{(\ell)}, \qquad \mathbf{W}_2^{(\ell)} = \mathbf{W}_b^{(\ell)} + \mathbf{A}_2^{(\ell)}, \tag{8}$$

where $\mathbf{C}^{(\ell)}$ denotes the unconstrained trainable update, which is parameterized in practice by LoRA adapters Hu et al. (2021), while $\mathbf{\Pi}_\perp^{(\ell)}$ is precomputed and frozen throughout training.

This design constrains the effective control update to the orthogonal complement of the dominant discrepancy subspace, thereby reducing its overlap with the principal feature directions associated with the bidirectional-to-causal transition, as illustrated in Fig. 2(c). Please refer to the next section for the theoretical insights behind this strategy.

4 Theoretical Analysis

In this section, we provide theoretical insights for the proposed orthogonal decoupled training strategy. Our analysis takes inspiration from recent study on model merging Cheng et al. (2025) and composable diffusion models Liu et al. (2022).

Theorem 1 (Informal Version: Feature-Level Decoupling)**.** *Let ΔW_1 and ΔW_2 denote two functionality-specific parameter updates. Suppose the dominant singular subspace of ΔW_1 is removed from ΔW_2 via*

$$\widetilde{\Delta W}_2 = \Delta W_2(I - P_k), \tag{9}$$

where P_k projects onto the top singular directions of ΔW_1.

Then the interference induced by the second functionality on the first functionality is upper bounded by the residual tail energy outside the dominant singular subspace of ΔW_1. In particular, if the first functionality is approximately low-rank, the induced interference becomes negligible.

Meanwhile, the capacity reduction of the second functionality depends only on its overlap with the removed singular subspace.

Theorem 2 (Informal Version: Orthogonality Improves Output Additivity)**.** *Consider a nonlinear model $f(W)$ and two functionality-specific updates ΔW_1 and $\widetilde{\Delta W}_2$. The deviation from ideal additive composition is measured by*

$$f(W + \Delta W_1 + \widetilde{\Delta W}_2) - f(W + \Delta W_1) - f(W + \widetilde{\Delta W}_2) + f(W). \tag{10}$$

If the dominant singular subspace of ΔW_1 is removed from ΔW_2, then the higher-order interaction between the two functionalities becomes bounded by the residual tail singular values of ΔW_1.

Consequently, orthogonal subspace projection suppresses nonlinear coupling between functionalities and improves output compositionality.

The formal statement and proof are provided in Appendix B.

Intuitively, the above analysis suggests that enforcing orthogonality between the optimization directions of video controllability and model causality preserves the feature components required by both objectives, thereby yielding a bounded output error due to the inherent compositionality of diffusion model outputs.

5 Experiments

5.1 Experimental Settings

Training Settings. We implement SVEET with Wan2.1-1.3B-VACE Jiang et al. (2025) as the bidirectional backbone and chunk-wise Causal Forcing Zhu et al. (2026) as the default streaming backbone. The VACE control branch is optimized using LoRA adapters Hu et al. (2021) with rank 128, with all pretrained backbone parameters kept frozen. We adopt the AdamW optimizer Loshchilov & Hutter (2017) with a learning rate of 1×10^{-4} and weight decay of 10^{-2}. For ODT, the layer-wise projector $\Pi_\perp$ retains 80% of the cumulative spectral energy at each DiT block. Each task is trained for 10 epochs on a single NVIDIA A100 (80GB) with batch size 1, using 81-frame video clips at 480×832 resolution. We mainly evaluate three representative editing tasks: style transfer, video inpainting, and depth-to-video generation. Further training details are provided in Appendix C.1.

Dataset Setup. For training, we construct task-specific datasets by sampling videos from existing datasets. Specifically, style transfer data is sampled from Ditto Bai et al. (2025b), while videos for inpainting and depth-to-video generation are sampled from VPData Bian et al. (2025). For the depth-to-video task, depth conditions are further extracted using Video-Depth-Anything Chen et al. (2025). Each task contains approximately 6K–13K training video samples. For evaluation, we construct held-out test sets for each task. Specifically, we randomly sample 120 video pairs from Ditto Bai et al. (2025b) for style transfer, and 80 samples from VPData Bian et al. (2025) for both inpainting and depth-to-video tasks. For depth-to-video evaluation, depth maps are generated using the same depth estimation pipeline. More dataset details are provided in Appendix C.2.

Baselines. We compare our method against two categories of baselines. For existing approaches, we consider (i) SDEdit+CF, a training-free baseline that adapts SDEdit Meng et al. (2021) to Causal Forcing Zhu et al. (2026), (ii) StreamDiffusionV2 Feng et al. (2025), (iii) Daydream+CF Fosdick (2026), (iv) LiveEdit Wang et al. (2026). As open-source real-time video editing models remain scarce, we further construct two controlled baselines that share our overall pipeline but differ in

Table 1: Quantitative results on three video editing tasks.

(a) Video Style Transfer

Method	VLM	Text Alignment	VBench Evaluation					
	Editing Accuracy ↑	CLIP ↑	Subject Consistency ↑	Background Consistency ↑	Temporal Flickering ↑	Motion Smoothness ↑	Aesthetic Quality ↑	Overall Consistency ↑
SDEdit+CF	3.1328	0.1412	0.8674	0.8949	0.9734	0.9840	0.5134	0.0916
SDV2	4.2297	0.1833	0.9011	0.9238	**0.9857**	0.9798	0.4650	0.1055
DayDream+CF	5.2321	0.2276	0.9147	0.9203	0.9775	0.9882	0.5049	0.0872
LiveEdit	5.8672	0.2279	0.9276	0.9219	0.9832	0.9892	0.5289	0.1098
3D VACE attn	7.0653	0.2105	0.9313	0.9175	0.9592	0.9809	0.5335	0.1046
Channel concat	1.8902	0.1414	0.7259	0.8848	0.9703	0.9763	0.3655	**0.1229**
Ours	**7.4322**	**0.2287**	**0.9447**	**0.9298**	0.9819	**0.9898**	**0.5550**	0.1123

(b) Video Inpainting

Method	VLM	Text Alignment	VBench Evaluation					
	Editing Accuracy ↑	CLIP ↑	Subject Consistency ↑	Background Consistency ↑	Temporal Flickering ↑	Motion Smoothness ↑	Aesthetic Quality ↑	Overall Consistency ↑
SDEdit+CF	6.7137	0.2816	0.9422	0.9363	0.9863	0.9909	0.5496	0.0928
SDV2	6.8580	0.2654	0.9609	**0.9551**	0.9836	**0.9945**	0.4867	0.0964
DayDream+CF	7.0115	0.2807	0.9579	0.9420	0.9880	0.9935	0.5434	0.0955
LiveEdit	7.7833	0.2803	0.9632	0.9476	0.9873	0.9933	**0.5498**	0.0962
3D VACE attn	8.0583	0.2882	0.9554	0.9426	0.9866	0.9933	0.5252	0.0947
Channel concat	6.8913	0.2690	0.9475	0.9369	0.9840	0.9933	0.4935	0.0926
Ours	**8.4667**	**0.2895**	**0.9653**	0.9478	**0.9886**	0.9936	0.5297	**0.1014**

(c) Depth-to-Video

Method	VLM	Text Alignment	VBench Evaluation					
	Editing Accuracy ↑	CLIP ↑	Subject Consistency ↑	Background Consistency ↑	Temporal Flickering ↑	Motion Smoothness ↑	Aesthetic Quality ↑	Overall Consistency ↑
SDEdit+CF	4.3673	0.2658	0.9239	0.9208	0.9920	0.9943	0.4854	0.0929
SDV2	5.0505	0.2275	0.9411	0.9513	**0.9961**	**0.9961**	0.4606	**0.1030**
DayDream+CF	7.1502	0.2892	0.9332	0.9239	0.9818	0.9884	0.5096	0.0811
LiveEdit	5.1892	0.2896	0.9613	0.9507	0.9958	0.9842	0.5273	0.0809
3D VACE attn	5.7484	0.2907	0.9720	0.9517	0.9855	0.9919	0.5466	0.0793
Channel concat	6.0830	0.2892	0.9332	0.9239	0.9818	0.9884	0.5096	0.0811
Ours	**8.4083**	**0.2928**	**0.9727**	**0.9521**	0.9892	0.9882	**0.5511**	0.0813

key design choices: (v) Full-Attn, which retains full spatiotemporal attention at control branch, and (vi) Channel-Concat Control, which replaces the control branch with channel-wise concatenation of source and noisy latents.

Evaluation Metrics. We evaluate edited videos mainly from three perspectives. (1) VLM Assessment: we use GPT-4o Hurst et al. (2024) to assess editing quality, including faithfulness to the instruction and visual coherence, and report average scores on a 1–10 Likert scale; (2) Text alignment: we assess this using CLIP-T Radford et al. (2021), computed as the cosine similarity between the editing prompt and sampled video frames encoded by CLIP ViT-L/14; (3) VBench Evaluation: we evaluate video quality with six VBench Huang et al. (2024) metrics relevant to our tasks including subject consistency, background consistency, temporal flickering, motion smoothness, aesthetic quality and overall consistency. In addition, we conduct a user study involving 10 participants, who are asked to score 15 generated videos in terms of editing correctness, structural preservation, and motion smoothness.

5.2 Main Results

Quantitative Comparison. Tab. 1 presents quantitative comparisons across the three editing tasks on our test sets. Overall, our method achieves the best average performance across most metrics and tasks, with particularly strong gains in VLM scores, where it consistently outperforms all baselines by a clear margin. Tab. 3 further reports the results of our human evaluation, showing that videos generated by ours are consistently preferred by participants over existing methods.

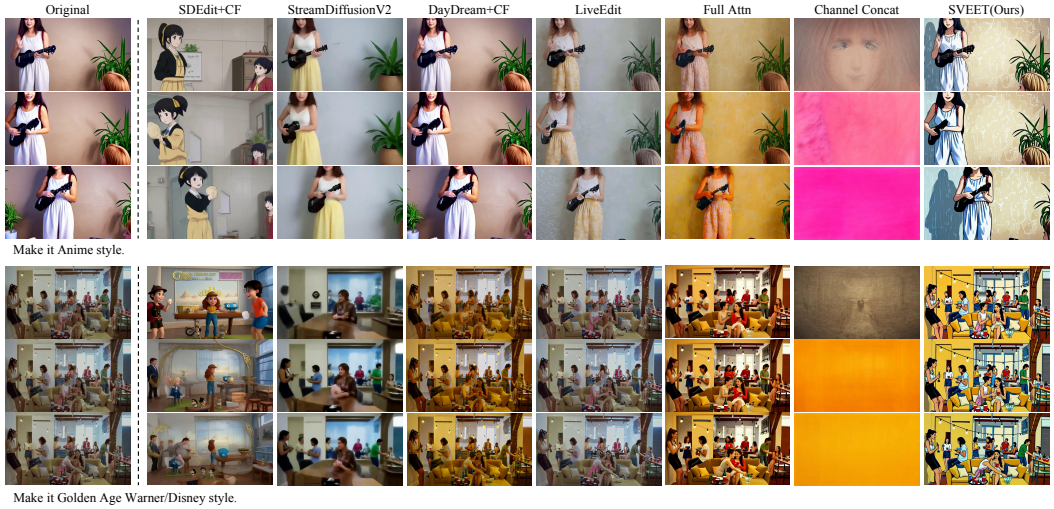


Figure 4: Qualitative comparison results on stylization.

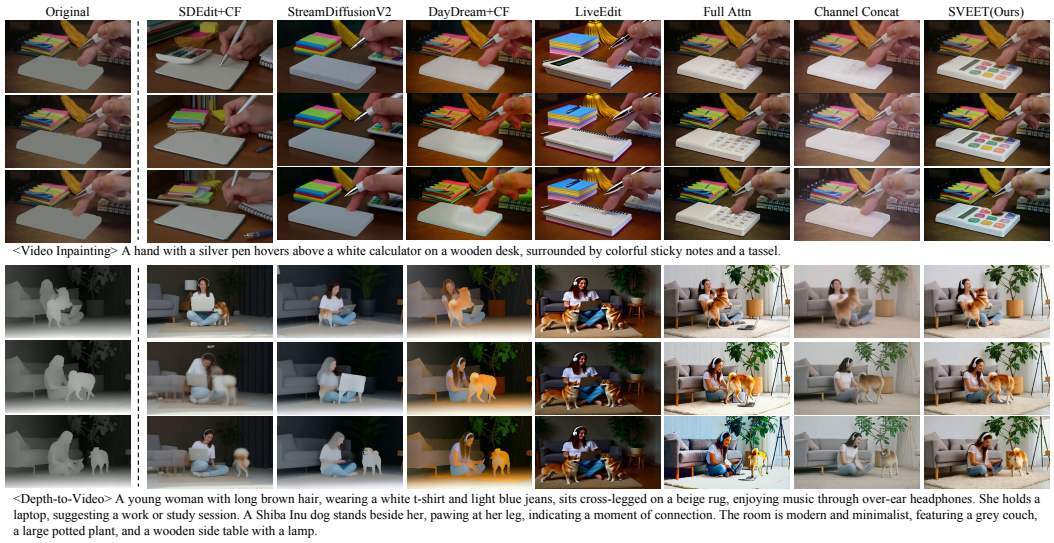


Figure 5: Qualitative comparison results on inpainting and depth-to-video generation.

Qualitative Comparison. We present qualitative results in Fig. 4 and Fig. 5. As shown in Fig. 4, our method achieves the best overall performance on the stylization task, consistently preserving source structure, faithfully transferring the target style, and maintaining strong temporal consistency. In contrast, SDEdit+CF Meng et al. (2021) and StreamDiffusionV2 Feng et al. (2025) significantly alter the original video content, while Channel Concat often fails to produce meaningful results. Daydream+CF and LiveEdit Fosdick (2026) struggle to generate the target style, and Full-Attn baseline shows weaker style intensity and inferior local detail alignment compared to our method. As shown in Fig. 5, on inpainting and depth-to-video tasks, our method also produces higher-quality outputs with noticeably better temporal stability. Baselines such as SDEdit+CF, StreamDiffusionV2, and Daydream+CF exhibit weaker generation quality, while LiveEdit, Channel Concat and Full-Attn variants consistently underperform in terms of visual fidelity and overall consistency. Additional transfer results on other streaming backbones are provided in Appendix D.2.

5.3 Ablation Studies

Ablations on transfer strategies. We first compare four controlled variants to disentangle the effects of temporal attention design and ODT: 3D/2D attention with or without ODT.

As shown in Tab. 2 and Fig. 6, 2D attention improves source preservation and streaming stability, while ODT further reduces transfer-induced appearance drift. Their combination achieves the best overall trade-off, supporting the complementarity of temporal independence and feature-space decoupling. We also compare two alternative transfer strategies. Inference-stage projection where post-hoc projection using the estimated mapping matrix causes substantial structural drift, while two-stage teacher forcing improves stability but still degrades local content and requires additional adaptation of the streaming backbone.

Table 2: Quantitative ablation results.

Settings	VLM ↑	Motion Smoothness ↑
3D w/o ODT	7.0653	0.9809
3D w/ ODT	7.3315	0.9856
2D w/o ODT	7.1898	0.9830
2D w/ ODT	**7.4322**	**0.9898**
Infer-stage Proj	6.4250	0.9848
Two-stage TF	7.0927	0.9879



Figure 6: Visual comparison of different transfer strategies.

Table 3: User study.

Method	Editing Correctness ↑	Structural Preservation ↑	Overall Smoothness ↑
SDEdit+CF	5.3867	2.8133	4.8467
SDV2	5.3267	2.9133	6.5800
DayDream+CF	7.8533	8.6200	7.9533
Ours	**9.0800**	**9.2067**	**8.1133**

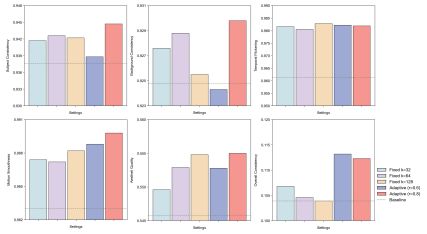


Figure 7: Effect of different rank choice.

Choice of subspace rank. We further study the effect of the subspace rank k used in the orthogonal projection. Since the discrepancy between bidirectional and streaming backbones varies across different DiT layers, using a fixed rank for all layers is suboptimal. Therefore, we adopt an adaptive strategy that determines k according to cumulative spectral energy. Here we compare fixed-rank settings ($k = 32, 64, 128$) with adaptive variants that retain a fixed proportion of energy ($r = 60\%, 80\%$). As shown in Fig. 7, fixed-rank settings generally underperform adaptive strategies, with small fixed ranks yield only marginal improvements over direct transfer. In contrast, adaptive rank selection leads to stronger overall gains, with $r = 80\%$ achieving comparatively better performance across most metrics.

6 Conclusions

We present SVEET, a high-quality streaming video editing framework trained solely on a pretrained bidirectional video diffusion model. We identify the key principles enabling bidirectional-to-streaming transfer without backbone retraining, and based on these insights, propose a temporally independent control branch together with an orthogonal decoupled training scheme to bridge the feature gap between bidirectional and autoregressive models for zero-shot streaming transfer. Extensive experiments demonstrate that SVEET achieves strong editing quality while maintaining real-time performance, providing a simple and effective paradigm for bridging offline and streaming video generation without costly retraining or distillation. In terms of limitations, our method still relies on the capability of the underlying bidirectional editing model, and extending this transfer paradigm to broader editing tasks and more heterogeneous backbones remains an direction for future work.

AI USE STATEMENT

Generative AI tools were used to assist with code editing and debugging, as well as language polishing and improving the readability of the manuscript. They were not used to formulate the core research ideas, design the proposed method, or interpret the experimental results. All AI-assisted code and text were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

As with other generative video technologies, the proposed method could potentially be misused to create misleading or manipulated visual content. We encourage responsible use of the technology and appropriate disclosure of AI-generated or AI-edited content.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed architecture and training procedure in Sec. 3.2 and Sec. 5. Implementation details, including model configurations, training hyperparameters, datasets, and evaluation protocols, are provided in Sec. 5.1. Additional implementation details and qualitative results are included in Appendix C and Appendix C.2. We have include the training and inference code in the supplementary material.

REFERENCES

Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10171–10180, 2025a.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025b.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

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.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025.

Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*, 2023.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. *arXiv preprint arXiv:2503.08099*, 2025.

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12736–12746, 2023.

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, et al. Streamdiffusionv2: A streaming system for dynamic and interactive video generation. *arXiv preprint arXiv:2511.07399*, 2025.

Ryan Fosdick. Adapting vace for real-time autoregressive video diffusion, 2026. URL `https://arxiv.org/abs/2602.14381`.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2568–2577, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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, 2021. URL `https://arxiv.org/abs/2106.09685`.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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*, pp. 21807–21818, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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*, pp. 17191–17202, 2025.

Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, Masayoshi Tomizuka, et al. Streamdiffusion: A pipeline-level solution for real-time interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12371–12380, 2025.

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.

Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27946–27956, 2025.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16818–16829, 2025a.

Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. Gas: Generative avatar synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12883–12893, 2025b.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yue Ma, Xiaodong Cun, Sen Liang, Jinbo Xing, Yingqing He, Chenyang Qi, Siran Chen, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9385–9395. IEEE, 2025.

Haitham Mahmoud and Raouf Abozariba. A systematic review on webrtc for potential applications and challenges beyond audio video streaming. *Multimedia Tools and Applications*, 84(6):2909–2946, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

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*, pp. 8748–8763. PmLR, 2021.

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*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4): 551–566, 1993.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7464–7473, 2019.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pp. 91–109. Springer, 2003.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xinyu Wang, Chongbo Zhao, Fangneng Zhan, and Yue Ma. Liveedit: Towards real-time diffusion-based streaming video editing. *arXiv preprint arXiv:2606.26740*, 2026.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation, 2024. URL https://arxiv.org/abs/2312.03641.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 2023. URL https://arxiv.org/abs/2212.11565.

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE transactions on visualization and computer graphics*, 31(2):1526–1541, 2024.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model, 2023. URL https://arxiv.org/abs/2311.16498.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025a.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. 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, 2025b.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation, 2026. URL https://arxiv.org/abs/2605.15141.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Empirical Verification of the Assumption

We further examine the feature-space assumptions underlying ODT. Fig. 8 shows the cumulative singular-value energy of the bidirectional-to-causal residual transformation $\mathbf{A}_1^{(\ell)}$ at representative DiT blocks. The spectra exhibit clear concentration: only a relatively small fraction of singular directions is required to capture most of the discrepancy energy. As summarized in Fig. 8(b), approximately 4%–21% of the directions are sufficient to explain 80% of the spectral energy across different layers. This observation supports our modeling of the bidirectional-to-causal shift with a compact dominant discrepancy subspace, while also motivating the layer-adaptive rank selection used in ODT.

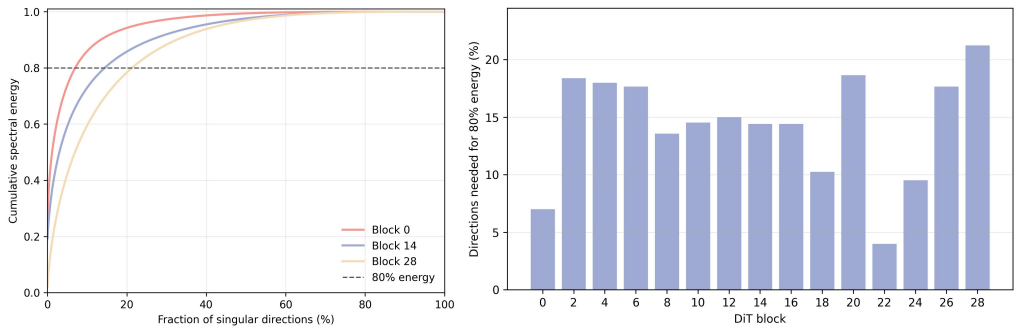


Figure 8: Empirical verification of the assumption.

B Theoretical Derivation

We present the full theoretical derivations corresponding to Sec. 4, including the problem setup, assumptions, and proofs, to establish the validity of the proposed orthogonal decoupled training method.

Setup. Consider a linear transformation

$$Y = XW, \tag{11}$$

where $X \in \mathbb{R}^{n\times d}$ denotes the input feature matrix and $W \in \mathbb{R}^{d\times m}$ denotes the model parameter.

Suppose two functionalities are represented by two parameter updates

$$\Delta W_1, \quad \Delta W_2. \tag{12}$$

To reduce interference between the two functionalities, we compute the truncated SVD of ΔW_1:

$$\Delta W_1 = U\Sigma V^\top, \tag{13}$$

and define $V_k \in \mathbb{R}^{d\times k}$ as the matrix containing the top-k right singular vectors. We further define the orthogonal projector

$$P_k = V_k V_k^\top. \tag{14}$$

The second update is constrained to lie in the orthogonal complement of the dominant subspace of ΔW_1:

$$\widetilde{\Delta W}_2 = \Delta W_2(I - P_k). \tag{15}$$

Assumption. Assume there exists a projection matrix P such that the first functionality depends only on the subspace $\mathrm{Im}(P)$, i.e.,

$$X = XP + X(I - P), \tag{16}$$

and

$$XP\Delta W_1 = X\Delta W_1. \tag{17}$$

Equivalently,

$$X(I - P)\Delta W_1 = 0. \tag{18}$$

Interference Definition. After jointly applying the two updates,

$$W' = W + \Delta W_1 + \widetilde{\Delta W}_2, \tag{19}$$

we define the induced interference on functionality 1 as

$$E_1 = XP\widetilde{\Delta W}_2, \tag{20}$$

and the removed component of functionality 2 caused by the orthogonality constraint as

$$E_2 = X\Delta W_2 P_k. \tag{21}$$

Theorem 1 (Orthogonal Subspace Decoupling)**.** *Under the above setup, the following statements hold.*

(1) Functional Interference Bound. *The interference induced on functionality 1 satisfies*

$$\|E_1\|_2 \le \sigma_{k+1}(XP)\,\|\Delta W_2\|_2, \tag{22}$$

where $\sigma_{k+1}(XP)$ denotes the $(k+1)$-th singular value of XP.

(2) Capacity Loss Bound. *The removed component of functionality 2 satisfies*

$$\|E_2\|_F^2 \le \|X\|_2^2 \|\Delta W_2 P_k\|_F^2. \tag{23}$$

Moreover,

$$\|\widetilde{\Delta W}_2\|_F^2 = \|\Delta W_2\|_F^2 - \|\Delta W_2 P_k\|_F^2. \tag{24}$$

Therefore, the removed energy is exactly characterized by the overlap between ΔW_2 and the dominant singular subspace of ΔW_1.

Proof. By definition,

$$E_1 = XP\Delta W_2(I - P_k). \tag{25}$$

Since P_k projects onto the top-k right singular subspace, the Eckart–Young theorem gives

$$\|XP(I - P_k)\|_2 = \sigma_{k+1}(XP). \tag{26}$$

Thus,

$$\begin{aligned} \|E_1\|_2 &= \|XP(I - P_k)\Delta W_2(I - P_k)\|_2 && (27)\\ &\le \|XP(I - P_k)\|_2 \|\Delta W_2\|_2 && (28)\\ &= \sigma_{k+1}(XP)\|\Delta W_2\|_2. && (29) \end{aligned}$$

For the second part,

$$E_2 = X\Delta W_2 P_k. \tag{30}$$

Applying the standard inequality

$$\|AB\|_F \le \|A\|_2 \|B\|_F, \tag{31}$$

we obtain

$$\|E_2\|_F^2 \le \|X\|_2^2 \|\Delta W_2 P_k\|_F^2. \tag{32}$$

Finally, since

$$\widetilde{\Delta W}_2 = \Delta W_2(I - P_k), \tag{33}$$

and

$$P_k(I - P_k) = 0, \tag{34}$$

the Pythagorean decomposition gives

$$\|\Delta W_2\|_F^2 = \|\Delta W_2 P_k\|_F^2 + \|\widetilde{\Delta W}_2\|_F^2. \tag{35}$$

This completes the proof. □

Discussion. The theorem reveals a clear trade-off between compositionality and functional capacity. If the first functionality is approximately low-rank, then

$$\sigma_{k+1}(XP) \approx 0, \tag{36}$$

implying that the interference induced by the second functionality is tightly bounded. Meanwhile, the capacity reduction of the second functionality depends only on its overlap with the dominant singular subspace of the first functionality.

Intuitively, Theorem B indicates that the compositional error in functionality 1 is upper-bounded by the singular values of the minor components, which are typically small in practice, while the capacity loss for functionality 2 is upper-bounded by those of the major components. Fortunately, the update components for functionality 2 are trainable in our setup. We can explicitly enforce the model to adapt to functionality 2 without affecting the major components of functionality 1, thereby ensuring minimal interference between the two.

Output Compositionality Analysis. We further analyze how orthogonal subspace projection improves the additivity of model outputs under nonlinear transformations.

Let

$$f : \mathbb{R}^{d\times m} \to \mathbb{R}^p \tag{37}$$

denote the nonlinear model mapping from the parameter space to the output space.

Consider two parameter updates

$$\Delta W_1, \quad \widetilde{\Delta W}_2 = \Delta W_2(I - P_k), \tag{38}$$

where $P_k = V_k V_k^\top$ is the projector onto the dominant right-singular subspace of ΔW_1.

We define the compositionality error as

$$E_{\text{comp}} = f(W + \Delta W_1 + \widetilde{\Delta W}_2) - f(W + \Delta W_1) - f(W + \widetilde{\Delta W}_2) + f(W). \tag{39}$$

This quantity measures the deviation from ideal additive composition.

Theorem 2 (Orthogonality-Induced Compositionality)**.** *Assume f is twice continuously differentiable in a neighborhood of W, and let*

$$H_W = \nabla_W^2 f(W) \tag{40}$$

denote the Hessian operator at W.

Then the compositionality error satisfies

$$E_{\text{comp}} = \langle \Delta W_1, H_W \widetilde{\Delta W}_2 \rangle + o\Big(\|\Delta W_1\|_F \|\widetilde{\Delta W}_2\|_F \Big). \tag{41}$$

Furthermore, if the Hessian matrix H_W is diagonal and

$$\widetilde{\Delta W}_2 = \Delta W_2(I - P_k), \tag{42}$$

then

$$\|\Delta W_1^\top \widetilde{\Delta W}_2\|_F \le \sigma_{k+1}(\Delta W_1)\, \|\Delta W_2\|_F, \tag{43}$$

which yields

$$\|E_{\text{comp}}\| \le \|H_W\|_2\, \sigma_{k+1}(\Delta W_1)\, \|\Delta W_2\|_F + o(\cdot). \tag{44}$$

Therefore, the compositionality error is upper bounded by the residual tail singular value outside the dominant singular subspace of the first functionality.

Proof. Using the second-order Taylor expansion around W,

$$f(W + \Delta W) = f(W) + \langle J_W, \Delta W \rangle + \frac{1}{2}\langle \Delta W, H_W \Delta W \rangle + o(\|\Delta W\|^2), \tag{45}$$

where $J_W = \nabla_W f(W)$.

Substituting

$$\Delta W = \Delta W_1 + \widetilde{\Delta W}_2, \tag{46}$$

we obtain

$$f(W + \Delta W_1 + \widetilde{\Delta W}_2) = f(W) + \langle J_W, \Delta W_1 \rangle + \langle J_W, \widetilde{\Delta W}_2 \rangle \tag{47}$$

$$+ \frac{1}{2}\langle \Delta W_1, H_W \Delta W_1 \rangle \tag{48}$$

$$+ \frac{1}{2}\langle \widetilde{\Delta W}_2, H_W \widetilde{\Delta W}_2 \rangle \tag{49}$$

$$+ \langle \Delta W_1, H_W \widetilde{\Delta W}_2 \rangle + o(\cdot). \tag{50}$$

Subtracting

$$f(W + \Delta W_1) + f(W + \widetilde{\Delta W}_2) - f(W), \tag{51}$$

all first-order and self-quadratic terms cancel, yielding

$$E_{\text{comp}} = \langle \Delta W_1, H_W \widetilde{\Delta W}_2 \rangle + o(\cdot). \tag{52}$$

Finally, if H_W is diagonal, since

$$\widetilde{\Delta W}_2 = \Delta W_2 (I - P_k), \tag{53}$$

and P_k projects onto the dominant right-singular subspace of ΔW_1, the residual overlap satisfies

$$\|\Delta W_1^\top \widetilde{\Delta W}_2\|_F \leq \sigma_{k+1}(\Delta W_1)\, \|\Delta W_2\|_F. \tag{54}$$

Substituting this bound completes the proof.

□

Intuitively, the Hessian matrix H_W characterizes output interference between the two functionalities. According to recent studies Liu et al. (2022), large-scale diffusion models inherently support functional composition, indicating the potential diagonality of H_W. Moreover, in our practical setup, functionality 1 is responsible for diffusion timestep modulation, while functionality 2 handles spatial control, further suggesting their conceptual independence.

C Additional implementation details

C.1 Training Details

Model Configuration. We build SVEET upon Wan2.1-1.3B-VACE Jiang et al. (2025) as the bidirectional editing backbone. Unless otherwise specified, chunk-wise Causal Forcing Zhu et al. (2026) is used as the default streaming backbone. The pretrained parameters of both backbones remain frozen throughout training, and only the VACE control branch is optimized. Specifically, we insert rank-128 LoRA adapters Hu et al. (2021) into the q, k, v, o, $\mathrm{ffn}.0$, and $\mathrm{ffn}.2$ linear layers of the control branch. The original spatiotemporal self-attention in the VACE branch is replaced with frame-wise 2D spatial attention, such that condition features are extracted independently for each source frame.

Orthogonal Projector Construction. For Orthogonal Decoupled Training (ODT), we first construct a small calibration set including over 500 samples and feed identical video–prompt pairs to the bidirectional and streaming backbones. At each DiT block ℓ, we collect the corresponding hidden features $\mathbf{X}_b^{(\ell)}$ and $\mathbf{X}_c^{(\ell)}$ and estimate the bidirectional-to-causal linear transformation using the ridge-regression formulation described in Sec. 4. We then perform SVD on the residual transformation $\mathbf{A}_1^{(\ell)}$. The rank k is selected independently for each layer such that the top-k singular directions capture 80% of the cumulative discrepancy energy. The resulting orthogonal projector $\mathbf{\Pi}_\perp^{(\ell)}$ is computed once before training and remains fixed thereafter. During optimization, the learnable LoRA update is projected onto this orthogonal complement as described in Eq. 8.

Training Protocol. All models are optimized using AdamW Loshchilov & Hutter (2017) with a learning rate of 5×10^{-5} and weight decay of 10^{-2}. We use a batch size of 1 and train each editing task for 10 epochs on a single NVIDIA A100 GPU with 80GB memory. Training videos contain 81 frames at a spatial resolution of 480×832. We train separate task-specific control branches for style transfer, video inpainting, and depth-to-video generation, while keeping the underlying bidirectional backbone unchanged.

Streaming Transfer and Inference. After training, the learned control branch is directly attached to the frozen streaming backbone without any additional fine-tuning or distillation. No parameters of the streaming model are updated during this transfer. Since the control branch uses temporally independent 2D attention, its conditioning features can be computed frame by frame without introducing an additional temporal KV cache. For experiments on alternative streaming backbones, we keep the trained SVEET control branch unchanged and replace only the streaming model used at inference time; further results are reported in Sec. D.1.

C.2 Dataset Details

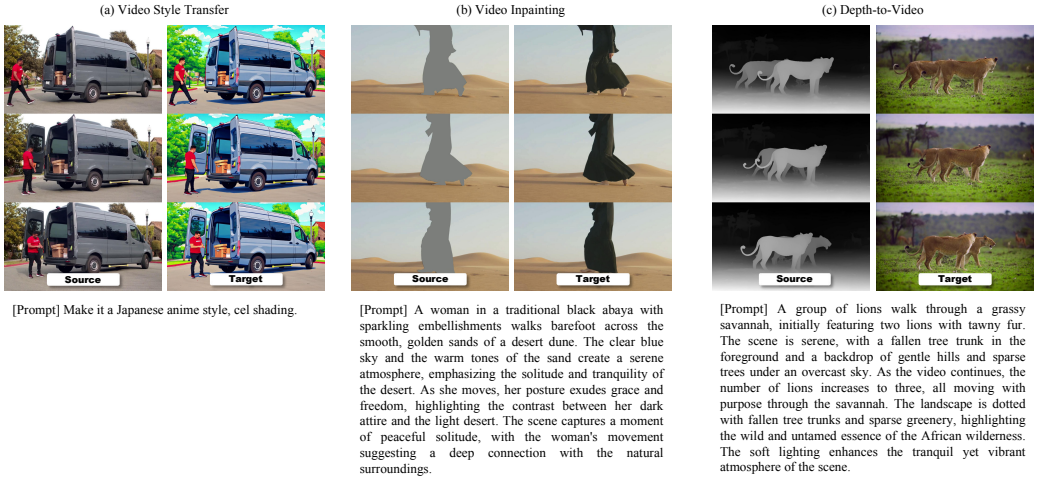


Figure 9: Representative training samples for three tasks.

We construct separate datasets for the three editing tasks considered in our experiments: style transfer, video inpainting, and depth-to-video generation. The training and evaluation splits are disjoint for all tasks.

Style Transfer. For video style transfer, we sample source videos from Ditto Bai et al. (2025b). Each training sample consists of a source video paired with a textual instruction specifying the target visual style. We use 26 distinct styles with 500 video samples per style, resulting in approximately 13K training pairs in total. For evaluation, we randomly select 120 held-out videos from Ditto that do not overlap with the training set.

Video Inpainting. For video inpainting, we sample source videos from VPData Bian et al. (2025), resulting in more than 6K training samples. Each sample contains the target video together with its masked conditioning input, where the inpainting masks follow the original mask annotations provided by VPData. We use 80 held-out VPData samples for evaluation.

Depth-to-Video Generation. For depth-to-video generation, we also sample videos from VPData Bian et al. (2025), obtaining approximately 7K training samples. For each video, frame-wise depth maps are extracted using Video-Depth-Anything Chen et al. (2025) and used as the structural condition for reconstructing the corresponding target video. The same depth extraction pipeline is applied to both training and evaluation data to ensure consistent conditioning. We randomly select 80 held-out videos for evaluation.

Preprocessing. All videos are converted to 81-frame clips at a spatial resolution of 480×832. Task-specific conditioning signals, including masked videos and depth maps, are temporally aligned

with their corresponding target frames. The resulting training sets contain approximately 13K samples for style transfer, over 6K samples for video inpainting, and 7K samples for depth-to-video generation. The evaluation sets contain 120 samples for style transfer and 80 samples each for video inpainting and depth-to-video generation. Representative training samples for the three tasks are shown in Fig. 9.

D Additional Experimental Results

D.1 Results on Other Streaming Models

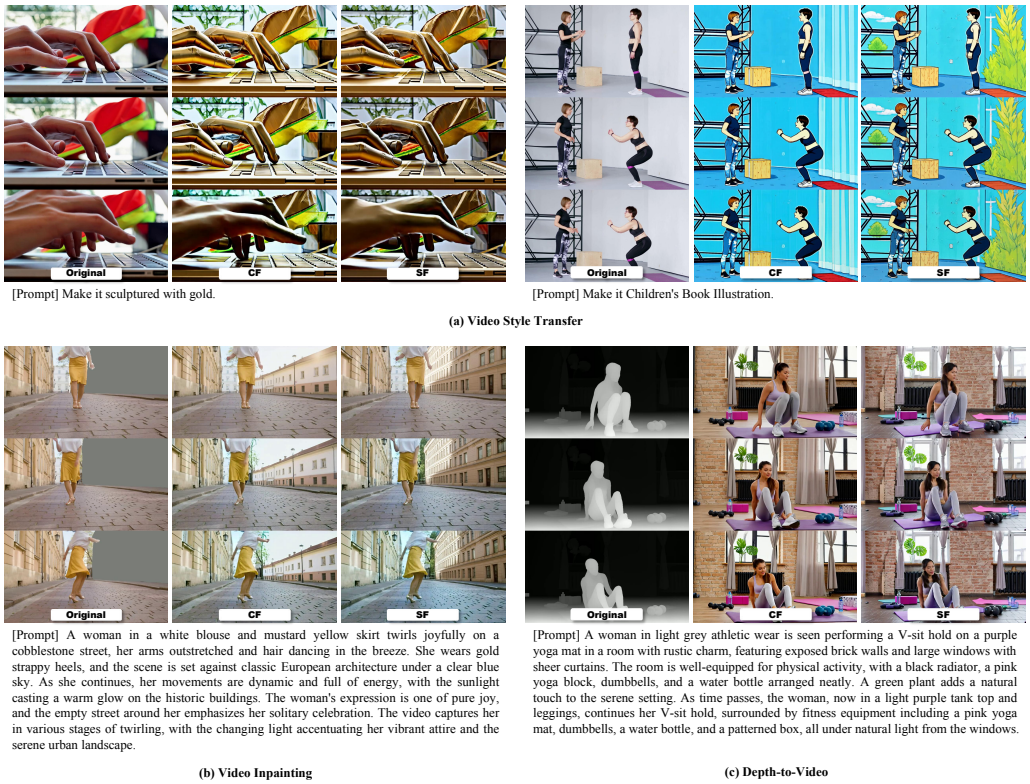
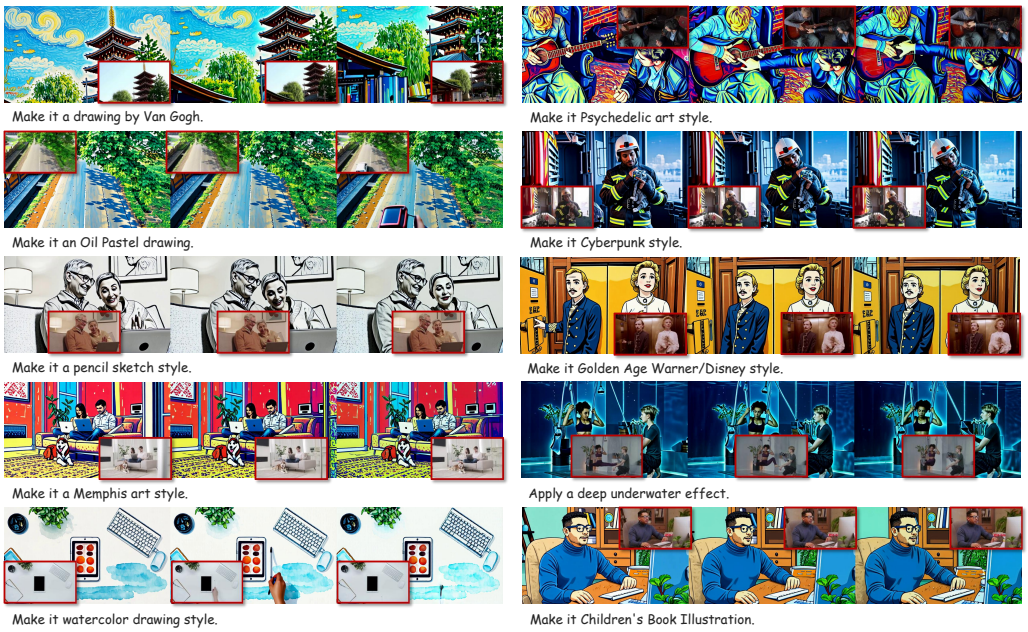


Figure 10: Transfer results on different streaming settings. Note that CF refers to Causal Forcing and SF refers to Self Forcing.

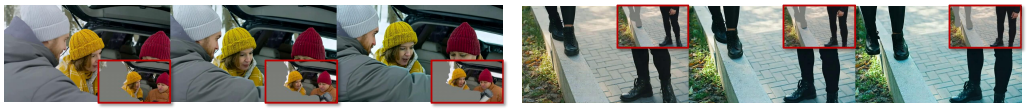
To further examine the transferability of our method across different streaming backbones, we additionally evaluate SVEET on Self Forcing Huang et al. (2025), in addition to the Causal Forcing backbone Zhu et al. (2026) used in the main experiments. Qualitative comparisons are shown in Fig. 10. Our method produces consistent editing effects on both Causal Forcing and Self Forcing, while preserving the main spatial structure and temporal dynamics of the source videos. These results indicate that the proposed bidirectional-to-streaming transfer strategy is not restricted to a single streaming backbone. This demonstrates the strong generalization capability and versatile applicability of our framework, revealing its great potential for universal video editing across diverse streaming-based generation architectures.

D.2 More Visual Results

We further present additional results generated by our method in Fig. 11, including diverse source videos, editing instructions, and control conditions. These examples demonstrate that SVEET can produce stable and visually consistent streaming editing results across a wide range of scenarios.



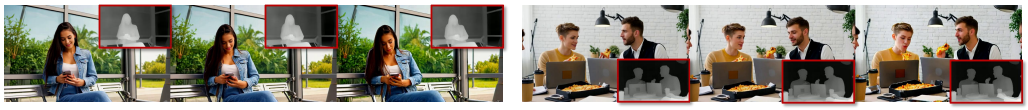
(a) Video Style Transfer



Three friends are enjoying a moment of camaraderie in a snowy setting, with two women and a man engaging in conversation by an open car trunk. They are dressed in warm winter attire, including mustard yellow beanie hats, a grey coat, and a brown jacket, indicating a chilly day. The man pours hot chocolate from a thermos into a cup for one of the women, who smiles playfully at him. Another woman watches with a smile. They are surrounded by a serene, snowy landscape, suggesting a peaceful outdoor activity or a break from driving.

Two individuals are seen standing on a concrete curb in an urban setting, both dressed in black combat boots with lace-up fronts and metal eyelets. Initially, one person is slightly ahead, wearing black leather pants, while the other, in a denim jacket, follows closely. As time passes, their attire remains consistent, with the addition of a white shirt under the jacket. The scene is set against a backdrop of a brick-paved pathway, suggesting a cooler season. The presence of greenery and the soft glow of natural light indicate a serene, possibly early morning or late afternoon, atmosphere.

(b) Video Inpainting



A young woman with long, dark hair is seated on a modern metallic bench at a bus stop, engrossed in her smartphone. She is dressed in a denim jacket and blue jeans, with a white top underneath. The scene is set against a backdrop of a clear glass panel reflecting a sunny day and lush greenery, suggesting a warm, possibly tropical climate. The woman remains focused on her phone throughout, with no other individuals or distractions around her, highlighting her solitary presence in the tranquil setting.

Four professionals are in a modern office, taking a casual break to eat pizza and converse. Initially, a man in a beige blazer and a woman in a yellow sweater are seen eating and interacting, with another man in a white shirt and black vest working on his laptop, and a fourth person partially visible. The table is scattered with laptops, a smartphone, and coffee cups. As time passes, the group continues their relaxed meeting, with the man in the beige blazer now eating pizza and talking to a woman in a yellow cardigan, while another man in a white shirt and black vest works on his laptop, and a woman in a yellow sweater types on her laptop. The office has a white brick wall, a black desk lamp, and a potted plant, creating a casual yet productive atmosphere.

(c) Depth-to-Video

Figure 11: More visualization results.