MSEditor: Toward Consistent Multi-Shot Video Editing

Kunyu Feng[1], Yue Ma[2], Bingyuan Wang[1], Yuefeng Wang[3], Zhiyuan Qin[4], Hao Cheng[5], Hao Li[4], Qifeng Chen[2], and Zeyu Wang[1,2]

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] Baidu Inc., Beijing, China
[4] Beijing Innovation Center of Humanoid Robotics, Beijing, China
[5] Tsinghua University, Beijing, China

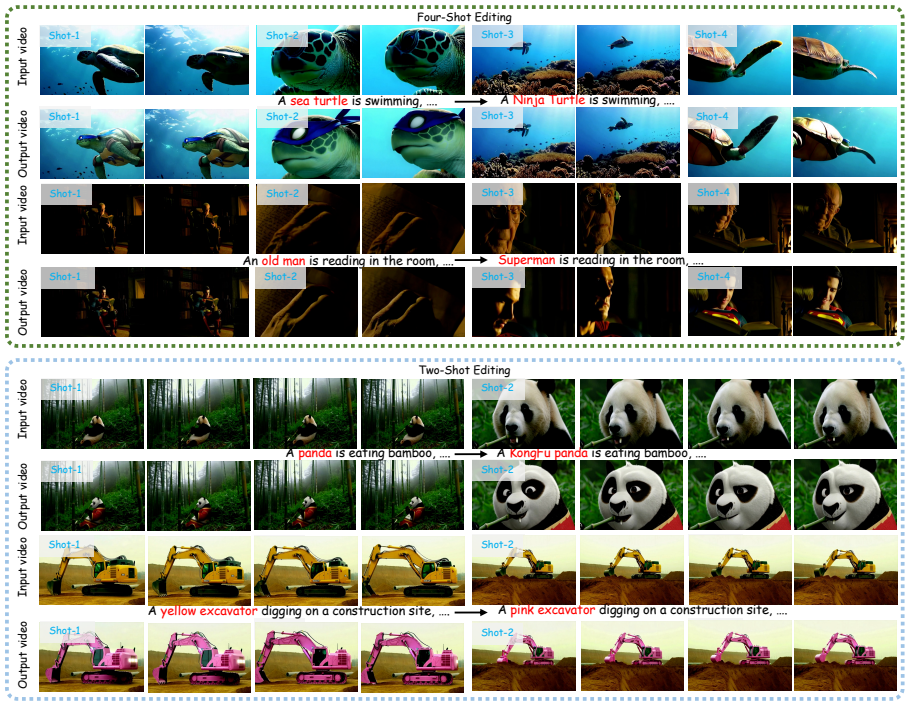

Fig. 1: Showcase of MSEditor. Given a multi-shot input video, our method produces consistent edits across shots with diverse scene compositions, maintaining appearance, motion, and cross-shot consistency.

Abstract. In this paper, we tackle the problem of performing consistent, unified modifications to a multi-shot video sequence. This task is par-

ticularly challenging because multi-shot videos consist of discontinuous temporal segments that vary significantly in viewpoint, camera scale, and subject pose, leading to severe identity drift and cumulative error propagation. Achieving coherent edits requires establishing reliable cross-shot semantic awareness to maintain stable subject appearance and visual continuity across these disjointed boundaries. To address this, we propose MSEditor, the first framework designed specifically for consistent multi-shot video editing. To overcome the scarcity of high-quality multi-shot training data, we repurpose existing multi-view video datasets to provide robust cross-shot supervision. Architecturally, we introduce a Supervisory Adapter that injects this cross-shot information into the diffusion backbone, enabling the model to learn identity-consistent representations. Furthermore, to effectively mitigate cumulative errors and ensure long-range temporal coherence, we design a Cross-Shot Packing strategy that dynamically aggregates information from semantically related shots within the self-attention window. Extensive experiments demonstrate that MSEditor significantly outperforms existing methods on our curated multi-shot video editing benchmark in terms of identity preservation, temporal stability, and overall visual quality.

Keywords: Multi-shot video editing · diffusion model · edit consistency

1 Introduction

Multi-shot video editing aims to achieve consistent and unified content modification across a sequence of discontinuous temporal segments (i.e., shots) assembled to convey a unified narrative. Unlike conventional single-shot video editing [5, 26, 34, 42, 48, 68, 70, 71], which focuses on generating plausible modifications within a single continuous temporal flow, this task requires maintaining editing coherence in both appearance and structure across multiple shots. This capability is crucial for professional applications such as filmmaking, narrative storytelling, and commercial advertising, where visual content is inherently constructed from sequences of multiple diverse shots to create cinematic rhythm and convey complex storylines. Consequently, extending generative editing capabilities from single-shot to multi-shot scenarios is not merely an incremental step, but a crucial leap toward practical, professional-grade video creation.

While coherent multi-shot editing is highly desirable, ensuring semantic consistency across diverse shots remains a formidable challenge. These different shots encompass abrupt transitions in both spatial and temporal dimensions, including changes in viewpoints, camera scales (e.g., from a wide establishing shot to a close-up), and subject poses. Achieving high-level semantic consistency across these narratively linked but geometrically disjointed shots presents significant technical hurdles. Previous editing approaches typically operate on a single-shot basis [19, 37]. When applied to multi-shot sequences, they inherently struggle, suffering from severe identity drift and cumulative errors across temporal boundaries. Specifically, extending a single-shot video editing framework to multi-shot scenarios typically encounters two major bottlenecks:

(1) The lack of multi-shot video datasets. A major bottleneck lies in the scarcity of large-scale, high-quality multi-shot video datasets. Constructing such datasets is extremely challenging, as it requires collecting sequences that preserve consistent subject identity across diverse scenes, lighting conditions, and transitions. Due to the absence of publicly available datasets that support robust multi-shot training, current diffusion-based video editing frameworks are often limited to single-shot settings and rely heavily on short-term temporal correlations. This restricts their generalization ability and hinders the development of multi-shot video editing.

(2) Cumulative error propagation and cross-shot inconsistency. Within a localized temporal window (e.g., a short sequence containing two shots), previous editing frameworks [19, 37] often suffer from cumulative error propagation. Small editing artifacts or subtle identity drifts emerging in earlier frames tend to amplify recursively as the video progresses. When crossing shot boundaries, these accumulated deviations manifest as noticeable inconsistencies in the subject's appearance, style, and structural integrity. For longer videos that exceed the model's temporal context (e.g., 81 frames for Wan2.1 [69]), current paradigms [3, 26] necessitate partitioning the sequence into independent temporal chunks for separate inference, as illustrated in Fig. 2. This fragmented processing fundamentally lacks cross-shot semantic awareness, causing severe subject identity drift between disjoint chunks. Furthermore, this approach is computationally inefficient, requiring multiple redundant inference passes that significantly increase the post-production overhead.

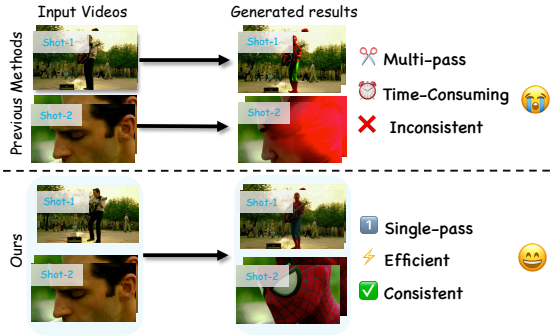

Fig. 2: Comparison with the previous single-shot editing framework. (Top) Existing single-shot editing baselines partition multi-shot sequences into independent temporal chunks. This fragmented inference leads to cumulative errors and cross-shot inconsistency, while being computationally time-consuming due to multiple inference passes. (Bottom) Our MSEditor performs a single inference pass, achieving high-fidelity consistency and superior efficiency.

To tackle these challenges, we propose **MSEditor**, a novel multi-shot video editing framework for coherent editing across diverse shots. To address the lack of multi-shot training data, we repurpose existing multi-view video datasets as a proxy for multi-shot supervision. Although they are originally designed for view synthesis, these datasets inherently contain sequences that capture the same subject under varying shots and scene conditions, providing valuable inter-

shot correspondence for training. Specifically, we extract synchronized RGB sequences along with their associated depth maps, from which we compute accurate and temporally stable subject masks through projection-based consistency. This avoids the segmentation failures of conventional models such as SAM2 [60], which often misidentify background figures in complex scenes. Leveraging these structured multi-view signals, we introduce a Supervisory Adapter that injects cross-shot supervision into the backbone, enabling the model to learn subject relationships and maintain appearance consistency across shots.

To further reduce error accumulation and improve cross-shot coherence, we introduce a Cross-Shot Packing Strategy. Our method dynamically groups semantically related shots and processes them jointly within the self-attention window. This enables effective cross-shot feature interaction during training, facilitating subject identity preservation and ensuring consistent edits across diverse shots.

In summary, our contributions are as follows:

- We introduce **MSEditor**, the first framework specifically designed for the multi-shot video editing task, achieving robust identity preservation and coherent editing across different shots.
- We propose a Supervisory Adapter that leverages auxiliary priors to inject cross-shot information into the backbone, enabling the model to learn identity-consistent representations for multi-shot video editing.
- We design a Cross-Shot Packing Strategy that dynamically aligns features across shots and mitigates cumulative error propagation, ensuring long-range temporal and appearance consistency.
- Extensive experiments demonstrate the state-of-the-art performance of our method on multi-shot editing tasks, showing superior temporal stability and identity consistency over existing methods.

2 Related Work

In this section, we review existing literature on single-shot video editing, multi-shot video generation, and long video generation.

Single-Shot Video Editing. In recent years, video diffusion models have reshaped single-shot video editing [67]. UNet-based approaches evolved from temporal mechanisms [8, 11, 14] to training-free feature propagation via inter-frame correspondences [19, 28, 64, 74]. Other advances include attention manipulation for coherence [37, 41, 47, 50, 57], adding spatiotemporal controls to image models [7, 72], and repurposing video models for editing [52, 77]. DiT-based methods exploit Diffusion Transformers [46, 56, 66] for scalability [3, 83], with innovations in query-key manipulation [38, 63, 65, 75, 82], dataset construction [73], unified in-context editing [15, 16, 26, 32, 78], and applications in autonomous driving [25], inpainting [44], and controllable video editing [17, 18, 31, 34, 42, 45, 49, 76]. However, these methods are designed for single shots and lack mechanisms for subject consistency across multiple shots with varying viewpoints.

Multi-Shot Video Generation. Recently, multi-shot video generation has advanced via camera-controlled and camera-free approaches [43]. Camera-controlled methods model camera parameters for multi-view synchronization [2], re-rendering [1], and pose control [23]. Camera-free methods focus on narrative coherence: transition tokens for shot control [27], Long Context Tuning to expand attention [20], HoloCine for long-range consistency [51], and SKALD for shot assembly [40]. Specialized datasets include AnimeShooter [59], TalkCuts [9], and Shot2story20k [21]. Despite the progress, these methods focus on generation from scratch rather than editing, and often require auxiliary geometric inputs, limiting practical use.

Long Video Generation. Long video generation focuses on coherence over extended sequences [29,36]. Early strategies used hierarchical latent diffusion [24] and parallel generation [79]. Later works enabled flexible frame sampling [22], sampling correction [12], slice-based editing [13], transition models [10], and noise rescheduling [58]. Recently, Mixture of Contexts (MoC) [6] uses sparse attention for near-linear scaling and minute-long coherence. However, these methods do not address the unique challenges of multi-shot editing, like cross-shot feature alignment and identity preservation across viewpoint changes.

3 Method

Our objective is to achieve temporally consistent and high-fidelity video editing across multiple discontinuous shots. Sec. 3.1 elaborates on our data construction and annotation approach. Sec. 3.2 presents our comprehensive framework for multi-shot video editing, which integrates a Supervisory Adapter for multi-shot information injection, a Cross-Shot Packing Strategy, and a Sparse Cross Attention Module for cross-shot consistency preservation.

3.1 Data Annotation and Collection

The development of multi-shot video editing techniques faces a fundamental challenge: the scarcity of high-quality datasets capturing sequential, diverse shots of the same subject. Manual collection of such data is prohibitively labor-intensive, significantly constraining model training and generalization capabilities.

To address this limitation, we repurpose existing multi-view video datasets [2, 30] as an effective proxy. These datasets provide synchronized sequences of identical subjects from multiple viewpoints, containing rich inter-view correspondences that are instrumental for learning cross-shot consistency. However, adapting these datasets for multi-shot editing requires careful consideration. Unlike generation tasks, video editing demands precise subject-scene disentanglement to ensure accurate editing. As illustrated in Fig. 4(a), direct application of off-the-shelf object segmentation models (e.g., SAM2 [60], Grounded-SAM [61]) often fails in complex scenarios where background elements contain distracting content such as paintings or incidental persons.

To overcome this limitation, we develop a robust annotation method depicted in Fig. 4(b). We employ double-reprojection techniques [80] to generate reliable

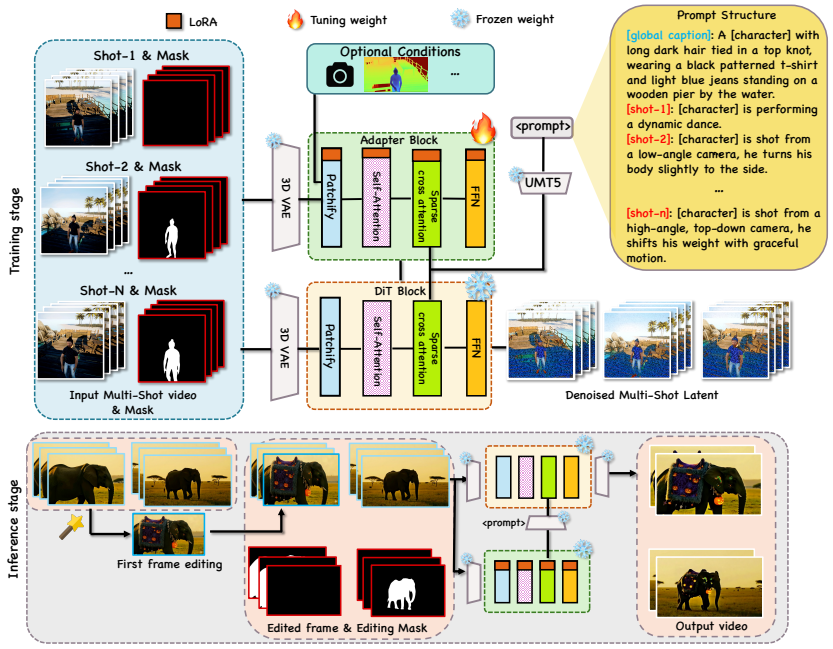

Fig. 3: Overview of our method. ***Top:*** Given an input multi-shot video and corresponding editing masks, our method first trains a Supervisory Adapter to inject multi-shot information into the diffusion backbone. We also introduce two key strategies: Cross-Shot Packing and Sparse Cross-Attention to keep the consistency during training. ***Bottom:*** During inference, editing is performed on the first frame and produces coherent and identity-preserving multi-shot video output.

depth maps and view-consistent segmentation masks. These annotations provide explicit spatial and structural supervision during training, enabling precise subject isolation. To provide fine-grained semantic guidance, we implement an automated structured prompting pipeline using Qwen3-VL [4]. For each multi-shot sequence, we generate a comprehensive global caption that describes the overarching subject identity and scene context. Additionally, we generate shot-specific descriptions for each viewpoint to capture unique camera angles and motion characteristics. This hierarchical annotation enables the model to learn both global semantic alignment and shot-level temporal dynamics.

The training dataset consists of 3,400 video sequences. For each video, there are synchronized RGB sequences from 10 different viewpoints $\{V^{(k)}\}_{k=1}^{10}$ paired with their corresponding segmentation masks $\{M^{(k)}\}_{k=1}^{10}$. The dataset also includes constructed depth maps $\{D^{(k)}\}_{k=1}^{10}$ and camera parameters $\{\text{cam}^{(k)}\}_{k=1}^{10}$. Each sequence is associated with a structured prompt:

$$\text{Prompt} = \{\text{Caption}_{\text{Global}}, \text{Caption}_{\text{Shot-1}}, \text{Caption}_{\text{Shot-2}}, \ldots, \text{Caption}_{\text{Shot-10}}\}. \tag{1}$$

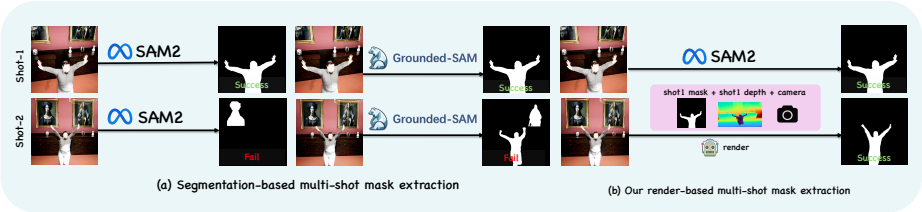

Fig. 4: Comparison of multi-shot mask extraction pipeline for training data construction. (a) Using segmentation models (e.g., SAM2 [60], Grounded-SAM [61]) naively on the per-shot video often results in inconsistent results. The model may successfully segment the subject in one shot but fail in another due to a complex background. (b) Our proposed method bypasses this limitation by leveraging the multi-view data's camera and depth information. We use a complete mask from one shot and re-project it into another's view, rendering an accurate mask. This process ensures high-quality, cross-shot, consistent annotations.

This comprehensive annotation format enables the Supervisory Adapter to learn implicit, cross-shot consistent representations, ensuring both appearance consistency and high-fidelity editing across diverse shots.

3.2 Multi-Shot Video Editing Framework

Our framework builds upon the pre-trained diffusion-based video generative backbone Wan2.1 [69], which has demonstrated strong performance in video synthesis within single-shot scenarios. To enable coherent editing across different shots, we introduce **MSEditor**, a novel architecture designed to preserve subject identity throughout multi-shot sequences. The overall framework is illustrated in Fig. 3.

Supervisory Adapter for Multi-Shot Information Injection. Although existing video editing frameworks exhibit strong editing capabilities in single-shot settings, they fail to maintain subject consistency across different shots due to the absence of multi-shot supervision. To enhance its multi-shot proficiency, we introduce a Supervisory Adapter that injects multi-shot features into the diffusion backbone during training.

As shown in Fig. 3, our adapter takes multi-view RGB videos and their corresponding segmentation masks as inputs. These multimodal signals guide the backbone to learn cross-shot consistent representations, enabling the model to preserve subject identity across discontinuous shots while maintaining high editing fidelity. Following the design principle of ControlNet [81], all adapter layers are initialized to zero to ensure stable convergence and prevent over-conditioning.

To leverage powerful prior knowledge and enhance user-friendliness control, we follow the recent video editing works [35,55] and adopt the first-frame editing paradigm that allows users to explicitly specify their editing intent on the initial frame. This reference-based approach significantly reduces the manual effort required for complex multi-shot sequences. Furthermore, to preserve the model's

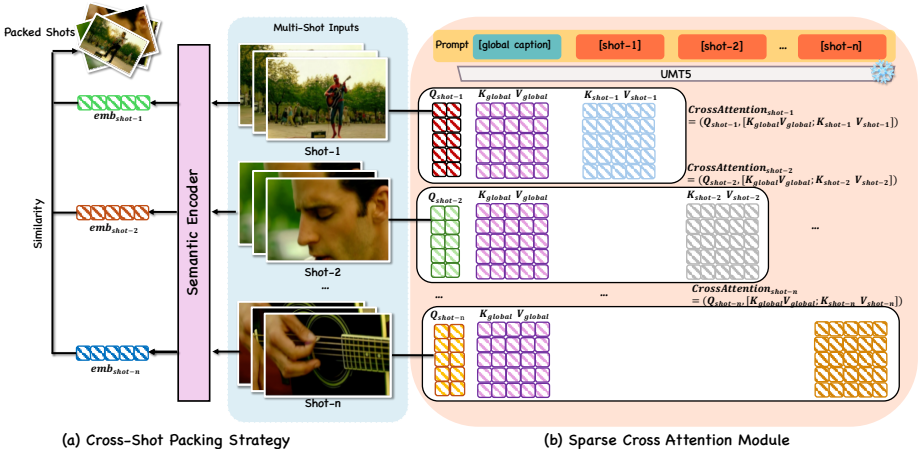

Fig. 5: Overview of the Cross-Shot Packing Strategy and Sparse Cross Attention Module. (a) Given multi-shot inputs, the Cross-Shot Packing Strategy addresses the challenge of maintaining consistency in multi-shot video editing. (b) The Sparse Cross Attention Module integrates textual guidance. Visual queries from each shot (e.g., $Q_{\text{shot-1}}$) sparsely attend only to the concatenation of the shared global text embeddings $(K_{\text{global}}, V_{\text{global}})$ and their corresponding shot-specific embeddings $(K_{\text{shot-1}}, V_{\text{shot-1}})$.

first-frame editing capability, we adopt a content editing strategy inspired by prior work [53]. Specifically, given a reference video

$$\mathbf{V} = [\mathbf{I}_0, \ldots, \mathbf{I}_{N-1}] \in \mathbb{R}^{N\times 3\times H\times W}, \tag{2}$$

where N denotes the number of frames, and H and W represent height and width respectively, we construct multi-shot video–mask pairs $\{(\mathbf{V}^{(k)}, \mathbf{M}^{(k)})\}_{k=1}^{K}$ corresponding to K different shots.

During the training stage, the mask of the first frame $\mathbf{M}_0^{(0)}$ is set to zero for the first shot, establishing the first frame as the guidance during the video generation, while subsequent masks define the editable regions for propagation through the sequence. During inference, users modify the first frame, and the changes are propagated to the following frames. Overall, the training objective minimizes the MSE loss between predicted and ground-truth frames across all video–mask pairs:

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{V}_{\text{pred}}^{(k)} - \mathbf{V}_{\text{gt}}^{(k)}\right\|_2^2. \tag{3}$$

This optimization enables the Supervisory Adapter to inject multi-shot priors into the backbone while ensuring the editing capability of the video foundation model, resulting in consistent multi-shot video editing.

Cross-Shot Packing Strategy for Consistent Object Preservation. While the Supervisory Adapter injects multi-shot information into our model, edit-

ing long videos with numerous shots remains challenging. In multi-shot scenarios, subjects may exhibit dramatic variations in scale, position, and appearance across shots. Minor inconsistencies in earlier shots can accumulate over time, leading to significant degradation of visual fidelity and structural coherence. However, in most training pipelines, multiple shots are typically concatenated along the batch dimension. Although this strategy improves training efficiency, it inherently isolates the temporal and semantic relations between different shots. Alternatively, concatenating all shots along the frame dimension would allow dense inter-shot interaction, but results in huge computational and memory costs.

To balance efficiency and consistency, we introduce a Cross-Shot Packing Strategy that reformulates self-attention computation by dynamically grouping semantically correlated shots into joint attention windows, as illustrated in Fig. 5(a). Formally, given a multi-shot video $\{V^{(k)}\}_{k=1}^{K}$, we randomly select a shot $V^{\text{anc}} = \{f_{\text{anc},t}\}_{t=1}^{T_{\text{anc}}}$ as the anchor shot, which retains its full temporal resolution to serve as the primary training target. For all other candidate shots $V^{(j)}$, we compute their semantic similarity to the anchor using DINOv2 [54] embeddings e_{anc} and e_j. We then select the top-M most correlated shots that exceed a threshold τ (e.g., 0.8). For each selected shot $V^{(j)}$, we extract a representative subset of n frames (e.g., the initial frames) to serve as identity frames:

$$\hat{V}^{(j)} = \mathcal{S}(V^{(j)}, n), \quad \text{where } n < F_j. \tag{4}$$

The final packed sequence $\tilde{V}^{(\text{anc})}$ is then formulated by concatenating the full anchor shot with the distilled reference frames:

$$\tilde{V}^{(\text{anc})} = \text{Concat}\left(V^{(\text{anc})}, \hat{V}^{(j_1)}, \ldots, \hat{V}^{(j_M)}\right). \tag{5}$$

To ensure the training remains robust under various GPU memory constraints, we implement a dynamic thresholding strategy. If the total frame count of $\tilde{V}^{(\text{anc})}$ exceeds the hardware's batch capacity, the system automatically increments τ (e.g., from 0.8 to 0.85) to filter out less relevant shots. This joint formulation enables the self-attention layers to perform feature interaction across semantically linked shots, facilitating a holistic understanding of subject identity across discontinuous temporal segments while maintaining computational tractability.

Textual Conditioning via Sparse Cross Attention. To integrate textual guidance, we adopt the Sparse Cross Attention mechanism from HoloCine [51] at both global and per-shot levels, as shown in Fig. 5(b). Given token embeddings of the textual prompt, we separate them into a global caption and multiple per-shot descriptions:

$$\text{Prompt} = \{\text{Caption}_{\text{Global}}, \text{Caption}_{\text{Shot-1}}, \text{Caption}_{\text{Shot-2}}, \ldots, \text{Caption}_{\text{Shot-K}}\}. \tag{6}$$

For each shot S_i, we compute cross-attention between its visual queries Q_i and the concatenated key–value pairs from global and corresponding per-shot

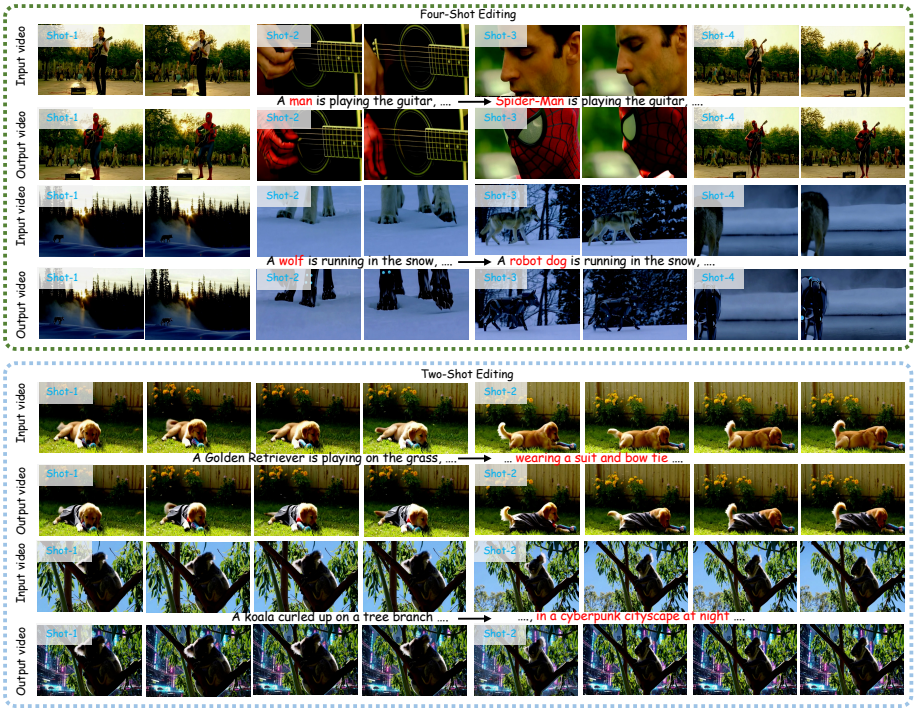

Fig. 6: Gallery of our proposed method, which displays four-shot and two-shot editing examples. These results demonstrate the effectiveness of our model in applying diverse semantic modifications while maintaining high visual fidelity and cross-shot consistency.

textual embeddings:

$$\text{Attn}_{\text{cross}}(Q_i) = \text{Attention}\left(Q_i, [K_{\text{global}}, K_i], [V_{\text{global}}, V_i]\right). \tag{7}$$

This formulation enables each shot to be guided by both global scene semantics and shot-specific textual context, thereby enhancing cross-shot consistency while preserving per-shot edit controllability.

4 Experiments

4.1 Experiment Setup

Datasets. Our model is trained on the multi-shot dataset constructed through the methodology detailed in Sec. 3.1. For evaluation, we curate a diverse test dataset comprising 600 videos collected from the internet, featuring a balanced distribution of 35% humans, 35% animals, and 30% objects. There is no overlap between our training and testing data, ensuring a rigorous assessment of

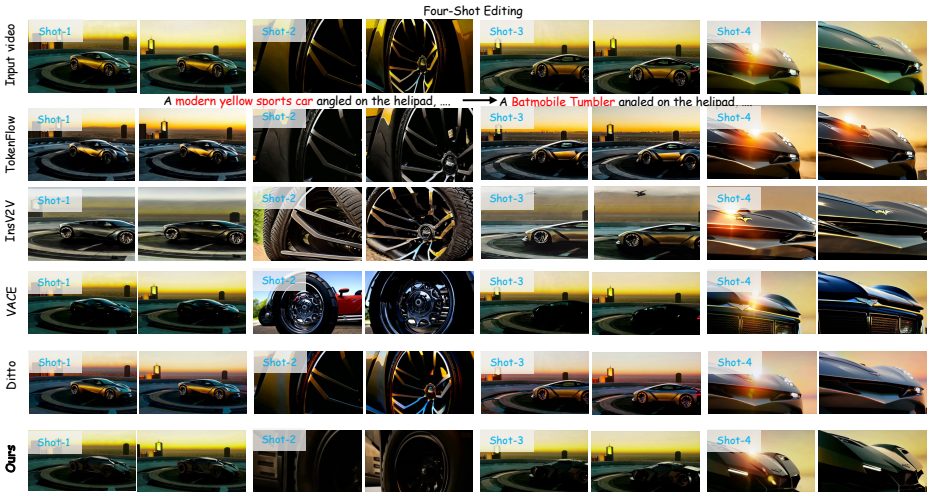

Fig. 7: Qualitative comparison results with the state-of-the-art methods. We compare our multi-shot approach against leading single-shot video editing baselines on a four-shot editing task. The results show that our method successfully executes the complex subject replacement, demonstrating superior cross-shot temporal consistency.

the model's generalization capabilities. The evaluation tasks encompass style transfer, object replacement, and attribute editing, with prompts automatically generated via Qwen3-VL [4]. More details of our test dataset are provided in the supplementary materials.

Implementation Details. We train our model based on Wan2.1-14B [69], a diffusion transformer-based video creation model. The optimization is performed using AdamW [39] with a weight decay of 0.01 and an initial learning rate of 1×10^{-4}. During the training process, we set the input resolution to 480×832 and a batch size of 8 on 8 NVIDIA A800 GPUs under PyTorch. During inference, we employ the flow-matching scheduler [33] with 50 sampling steps to generate high-quality and temporally coherent multi-shot videos. More evaluation metrics are provided in the supplementary material.

Table 1: Quantitative Results of Comparative Studies. **Red** and **Blue** denote the best and second best results, respectively.

Method	Video Quality	Inter-shot Consistency		Intra-shot Consistency		Semantic Consistency	
	Aesthetic Score ↑	Subject ↑	Background ↑	Subject ↑	Background ↑	Global ↑	Shot ↑
TokenFlow [19]	5.51	0.9181	0.9279	0.9029	0.9173	0.1302	0.1415
InsV2V [12]	4.89	0.9013	0.9307	0.8962	0.8941	0.1433	0.1137
VACE [26]	**5.91**	0.9237	**0.9372**	0.8805	**0.9365**	**0.1649**	**0.1738**
Ditto [3]	5.73	**0.9287**	0.9341	**0.9125**	0.9331	0.1514	0.1569
Ours	**6.05**	**0.9370**	**0.9447**	**0.9501**	**0.9462**	**0.1912**	**0.1886**

Table 2: Quantitative Results of Comparative Studies. For fair comparison, we train recent methods VACE and VideoPainter on our dataset. **Red** and **Blue** denote the best and second best results, respectively.

Method	Video Quality	Inter-shot Consistency		Intra-shot Consistency		Semantic Consistency	
	Aesthetic Score ↑	Subject ↑	Background ↑	Subject ↑	Background ↑	Global ↑	Shot ↑
VACE [26]	5.93	**0.9301**	0.9395	0.9057	**0.9382**	**0.1728**	**0.1753**
VideoPainter [5]	**6.01**	0.9237	**0.9396**	**0.9061**	0.9377	0.1687	0.1732
Ours	**6.05**	**0.9370**	**0.9447**	**0.9501**	**0.9462**	**0.1912**	**0.1886**

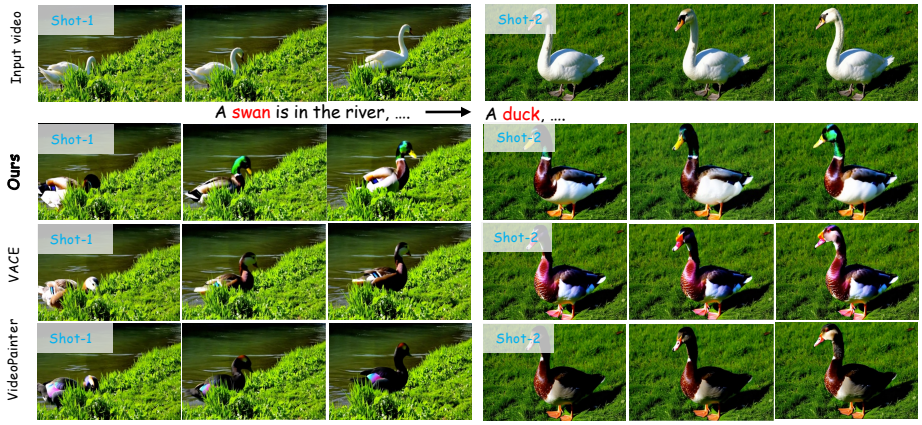

Fig. 8: More Qualitative Results after Training VACE and VideoPainter on Our Dataset. For fair comparison, we train VACE and VideoPainter on our dataset, and compare the performance on the two-shot editing task. The results demonstrate that our proposed method has superior performance.

4.2 Comparison with Baseline Methods

Qualitative Results. Due to the lack of existing methods for multi-shot video editing, we establish our baselines from the domain of single-shot video editing. To ensure a comprehensive comparison, we select methods spanning diverse architectures, namely UNet-based and DiT-based models. The chosen baselines include TokenFlow [19], InsV2V [12], Ditto [3], and VACE [26]. Furthermore, to ensure a fair comparison, we retrained VACE [26] and VideoPainter [5] on our curated multi-shot training dataset, enabling them to learn from identical data distributions. As shown in Fig. 7 and Fig. 8, we display the qualitative results of these methods on four-shot and two-shot video editing tasks. Specifically, previous methods such as InsV2V, TokenFlow, and Ditto fail to perform the required edits while maintaining coherence. Even after being trained on

our dataset, VACE and VideoPainter still struggle to maintain subject identity across discontinuous temporal boundaries, resulting in partial subject replacement or structural distortion. In contrast, our method demonstrates superior performance in editing fidelity, subject identity preservation, and temporal consistency. This indicates that the advantage of our approach stems not only from the training data but more importantly from our specialized architecture. Additional qualitative comparisons under more diverse settings are provided in the supplementary material.

Quantitative Results. For a quantitative study, we use the Aesthetic Score [62] to evaluate the quality of the edited video, and evaluate inter-shot consistency, intra-shot consistency, and semantic consistency following recent works [51]. As shown in Table 1, the quantitative results demonstrate that our proposed method outperforms the baseline across all metrics. To ensure a fair comparison, we retrained recent state-of-the-art methods, specifically VACE [26] and VideoPainter [5], using our own curated multi-view training dataset. As shown in Table 2, the quantitative results demonstrate that our proposed method outperforms all baselines across every metric. Notably, even after fine-tuning these baselines on the same data, our MSEditor still maintains a significant performance gap, which further validates the architectural superiority of our method. Furthermore, we provide a quantitative evaluation of our mask extraction accuracy and additional results for "w/o mask setting" and the shape-editing task in the supplementary material to further demonstrate our method's robustness. The results of our user study are also available in the supplementary material.

Table 3: Quantitative Results of the Ablation Study. **Red** and **Blue** denote the best and second best results, respectively.

Method	Video Quality	Inter-shot Consistency		Intra-shot Consistency		Semantic Consistency	
	Aesthetic Score ↑	Subject ↑	Background ↑	Subject ↑	Background ↑	Global ↑	Shot ↑
w/o First Frame editing	**6.01**	**0.9321**	**0.9357**	**0.9336**	**0.9305**	**0.1752**	**0.1729**
w/o Cross-Shot Packing strategy	5.83	0.8919	0.9132	0.8937	0.9008	0.1738	0.1521
w/o Sparse Cross attention	5.70	0.8931	0.9256	0.8651	0.9126	0.1691	0.1432
Ours	**6.05**	**0.9370**	**0.9447**	**0.9501**	**0.9462**	**0.1912**	**0.1886**

4.3 Ablation Study

Effectiveness of First Frame editing. To investigate the contribution of the First Frame editing strategy, we conduct a series of ablation studies on it. In this setting, the model relies solely on the text prompt and the learned cross-shot priors to generate the edited content. The experimental settings are the same during the ablation. As shown in Fig. 9 and Table 3, removing the first-frame anchor results in a slight decrease in consistency and aesthetic scores. However, this performance drop is relatively marginal compared to the impact of removing the other two strategies. This suggests that while first-frame editing

serves as a high-fidelity visual anchor and significantly enhances user control, the core ability to maintain cross-shot consistency is fundamentally driven by our designed architecture. The results further demonstrate that our framework has learned robust, implicit semantic correspondences across disjoint shots.

Effectiveness of Cross-Shot Packing strategy. We further assess the effectiveness of the proposed Cross-Shot Packing strategy in Fig. 9 and Table 3. Without the Cross-Shot Packing strategy, the model degenerates into the baseline training setting where multiple shots are concatenated along the batch dimension, and the edited video has the challenge of maintaining inter-shot consistency.

Effectiveness of Sparse Cross Attention. As shown in Table 3, removing Sparse Cross Attention leads to a clear degradation in inter-shot semantic consistency. In this setting, the model receives the same structured prompts, but we replace our designed module with a standard cross-attention mechanism. As shown in Table 3, this substitution leads to a marked degradation in inter-shot semantic consistency. Furthermore, in Fig. 9, we show the results when there is a lack of Sparse Cross Attention. This failure is primarily driven by severe semantic confusion. When processing the lengthy and structured concatenated prompts, a standard dense attention mechanism allows all video frames to attend globally to all text tokens. Consequently, the model struggles to disentangle the long context, often failing to accurately align specific shot-level textual instructions with their corresponding temporal segments.

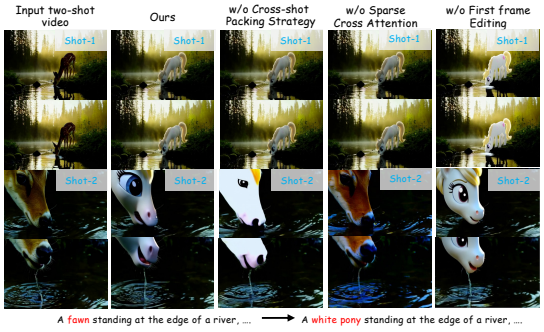

Fig. 9: Qualitative Results of the Ablation Study. We ablate the proposed modules in two-shot editing tasks. Although the first shot is successfully edited, (1) Without the Cross-Shot Packing Strategy, the result lacks inter-shot consistency, leading to visual artifacts (e.g., a distorted face in the close-up shot). (2) Without the Sparse Cross Attention, the model fails in the second shot, indicating a failure to process the per-shot semantic guidance. (3) Without the First Frame Editing, the model successfully transforms the target but exhibits subtle deviations in fine-grained details, highlighting its role as a high-fidelity visual anchor rather than a prerequisite for structural consistency.

5 Conclusion

We have proposed **MSEditor**, the first multi-shot video editing framework. Unlike prior single-shot methods that suffer from subject drift and temporal in-

consistency, our approach leverages structured multi-shot supervision during the training stage to endow the diffusion model with cross-shot awareness. Specifically, our Supervisory Adapter injects multi-shot information into the backbone, facilitating the structural coherence of the model. Additionally, we design the Cross-Shot Packing strategy to keep the consistency between different shots. Extensive experiments demonstrate that our approach substantially improves cross-shot alignment and subject fidelity compared to previous diffusion-based editing frameworks. We believe our framework opens promising avenues for future research in real-world, multi-shot storytelling applications.

6 Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 62502410), the Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515011138) and the Research Grants Council of HKSAR under grant number AoE/E-601/24-N.

References

1. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., Zhang, D.: ReCamMaster: Camera-Controlled Generative Rendering from a Single Video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
2. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints. In: Proceedings of the International Conference on Learning Representations (2025)
3. Bai, Q., Wang, Q., Ouyang, H., Yu, Y., Wang, H., Wang, W., Cheng, K.L., Ma, S., Zeng, Y., Liu, Z., Xu, Y., Shen, Y., Chen, Q.: Scaling Instruction-Based Video Editing with a High-Quality Synthetic Dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 37971–37981 (2026)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
5. Bian, Y., Zhang, Z., Ju, X., Cao, M., Xie, L., Shan, Y., Xu, Q.: VideoPainter: Any-length Video Inpainting and Editing with Plug-and-Play Context Control. In: ACM SIGGRAPH 2025 Conference Papers. pp. 1–12 (2025)
6. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., Agrawala, M., Jiang, L., Wetzstein, G.: Mixture of Contexts for Long Video Generation. In: Proceedings of the International Conference on Learning Representations (2026)

7. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2Video: Video Editing using Image Diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
8. Chai, W., Guo, X., Wang, G., Lu, Y.: StableVideo: Text-driven Consistency-aware Diffusion Video Editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23040–23050 (2023)
9. Chen, J., Wang, Z., Zeng, A., Fu, Y., Yu, X., Cen, S., Tanke, J., Chen, Y., Saito, K., Mitsufuji, Y., et al.: TalkCuts: A Large-Scale Dataset for Multi-Shot Human Speech Video Generation. Advances in Neural Information Processing Systems **38** (2025)
10. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction. In: The Twelfth International Conference on Learning Representations (2023)
11. Chen, Y., He, X., Ma, X., Ma, Y.: Contextflow: Training-free video object editing via adaptive context enrichment. arXiv preprint arXiv:2509.17818 (2025)
12. Cheng, J., Xiao, T., He, T.: Consistent Video-to-Video Transfer Using Synthetic Dataset. In: Proceedings of the International Conference on Learning Representations (2024)
13. Cohen, N., Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: SliceEdit: Zero-Shot Video Editing With Text-to-Image Diffusion Models Using Spatio-Temporal Slices. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 9109–9137 (2024)
14. Couairon, P., Rambour, C., Haugeard, J.E., Thome, N.: VidEdit: Zero-Shot and Spatially Aware Text-Driven Video Editing (2024), https://arxiv.org/abs/2306.08707
15. Deng, T., Chen, X., Chen, Y., Chen, Q., Xu, Y., Yang, L., Xu, L., Zhang, Y., Zhang, B., Huang, W., Wang, H.: Gaussiandwm: 3d gaussian driving world model for unified scene understanding and multi-modal generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10656–10667 (June 2026)
16. Deng, T., Chen, Y., Zhang, L., Yang, J., Yuan, S., Liu, J., Wang, D., Wang, H., Chen, W.: Compact 3d gaussian splatting for dense visual slam. arXiv preprint arXiv:2403.11247 (2024)
17. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion transformer for image editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2969–2977 (2025)
18. Gao, H., Tang, B., Song, Y., Fang, G., He, Z., Yang, J., Shou, M.Z.: Pai-studio: Cinematic video background replacement with camera-aware motion. arXiv preprint arXiv:2606.01399 (2026)
19. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In: Proceedings of the International Conference on Learning Representations (2024)
20. Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., Jiang, L.: Long Context Tuning for Video Generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17281–17291 (2025)
21. Han, M., Yang, L., Chang, X., Yao, L., Wang, H.: Shot2Story: A New Benchmark for Comprehensive Understanding of Multi-Shot Videos. In: Proceedings of the International Conference on Learning Representations (2025)

22. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible Diffusion Modeling of Long Videos. Advances in Neural Information Processing Systems **35**, 27953–27965 (2022)
23. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: CameraCtrl: Enabling Camera Control for Video Diffusion Models. In: Proceedings of the International Conference on Learning Representations (2025)
24. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent Video Diffusion Models for High-Fidelity Long Video Generation. arXiv preprint arXiv:2211.13221 (2022)
25. Jiang, J., Hong, G., Zhou, L., Ma, E., Hu, H., Zhou, X., Xiang, J., Liu, F., Yu, K., Sun, H., et al.: DiVE: DiT-based Video Generation with Enhanced Control. arXiv preprint arXiv:2409.01595 (2024)
26. 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)
27. Kara, O., Singh, K.K., Liu, F., Ceylan, D., Rehg, J.M., Hinz, T.: ShotAdapter: Text-to-Multi-Shot Video Generation with Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28405–28415 (2025)
28. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: AnyV2V: A Tuning-Free Framework for Any Video-to-Video Editing Tasks. Transactions on Machine Learning Research (2024)
29. Li, C., Huang, D., Lu, Z., Xiao, Y., Pei, Q., Bai, L.: A Survey on Long Video Generation: Challenges, Methods, and Prospects. arXiv preprint arXiv:2403.16407 (2024)
30. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3D Video Synthesis from Multi-View Video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5521–5531 (2022)
31. Liang, S., Guan, F., Zhang, Y., Li, X., Chen, Z.: Cot-edit: Let cot guide instruction video editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 37960–37970 (2026)
32. Liang, S., Wang, C., Guan, F., Yu, Z., Lu, Y., Wang, Y., Zhou, Y., Li, X., Chen, Z.: Spongebob: Sync-aware harmonious audio-visual generative editing. arXiv preprint arXiv:2605.25193 (2026)
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. In: Proceedings of the International Conference on Learning Representations (2023)
34. Liu, F.L., Fu, H., Wang, X., Ye, W., Wan, P., Zhang, D., Gao, L.: SketchVideo: Sketch-based Video Generation and Editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23379–23390 (2025)
35. Liu, F.L., Li, S.Y., Cao, Y.P., Fu, H., Gao, L.: Sketch3DVE: Sketch-based 3D-Aware Scene Video Editing. In: ACM SIGGRAPH 2025 Conference Papers. pp. 1–12 (2025)
36. Liu, J., Wang, X., Lin, Y., Wang, Z., Wang, P., Cai, P., Zhou, Q., Yan, Z., Yan, Z., Shi, Z., et al.: A survey on cache methods in diffusion models: Toward efficient multi-modal generation. arXiv preprint arXiv:2510.19755 (2025)
37. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-P2P: Video Editing with Cross-attention Control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)

38. Long, Z., Zheng, M., Feng, K., Zhang, X., Liu, H., Yang, H., Zhang, L., Chen, Q., Ma, Y.: Follow-your-shape: Shape-aware image editing via trajectory-guided region control. arXiv preprint arXiv:2508.08134 (2025)
39. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: Proceedings of the International Conference on Learning Representations (2019)
40. Lu, C.Y., Tanjim, M.M., Dasgupta, I., Sarkhel, S., Wu, G., Mitra, S., Chaterji, S.: SKALD: Learning-Based Shot Assembly for Coherent Multi-Shot Video Creation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17859–17868 (2025)
41. Lu, Z., Xiong, W., Ren, P., Jia, J.: Lighturban: Similarity based fine-grained instancing for lightweighting complex urban point clouds. Computer Graphics Forum **43** (11 2024). https://doi.org/10.1111/cgf.15238
42. Ma, Y., Cun, X., Liang, S., Xing, J., He, Y., Qi, C., Chen, S., Chen, Q.: 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)
43. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., Wang, B., Wang, Q., He, X., Wang, H., et al.: Controllable Video Generation: A Survey. arXiv preprint arXiv:2507.16869 (2025)
44. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. arXiv preprint arXiv:2506.04590 (2025)
45. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
46. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
47. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. arXiv preprint arXiv:2506.05207 (2025)
48. Ma, Y., Wang, X., Ma, Q., Wang, Q., Zheng, M., Yang, X., Li, H., Zhao, C., Ying, J., Yang, H., et al.: Group editing: Edit multiple images in one go. arXiv preprint arXiv:2603.22883 (2026)
49. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: Fastvmt: Eliminating redundancy in video motion transfer. arXiv preprint arXiv:2602.05551 (2026)
50. Ma, Y., Yan, Z., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., et al.: Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. arXiv preprint arXiv:2509.16630 (2025)
51. Meng, Y., Ouyang, H., Yu, Y., Wang, Q., Wang, W., Cheng, K.L., Wang, H., Ma, S., Li, Y., Chen, C., Zeng, Y., Zhu, X., Shen, Y., Qu, H.: HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 461–471 (2026)
52. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video Diffusion Models are General Video Editors. arXiv preprint arXiv:2302.01329 (2023)
53. Mou, C., Cao, M., Wang, X., Zhang, Z., Shan, Y., Zhang, J.: ReVideo: Remake a Video with Motion and Content Control. Advances in Neural Information Processing Systems **37**, 18481–18505 (2024)

54. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2024)
55. Ouyang, W., Dong, Y., Yang, L., Si, J., Pan, X.: I2VEdit: First-Frame-Guided Video Editing via Image-to-Video Diffusion Models. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
56. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
57. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023)
58. Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., Liu, Z.: FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling. In: Proceedings of the International Conference on Learning Representations (2024)
59. Qiu, L., Li, Y., Ge, Y., Ge, Y., Shan, Y., Liu, X.: AnimeShooter: A Multi-Shot Animation Dataset for Reference-Guided Video Generation. arXiv preprint arXiv:2506.03126 (2025)
60. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment Anything in Images and Videos. In: Proceedings of the International Conference on Learning Representations (2025)
61. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks (2024)
62. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
63. Shen, T., Huang, Z., Li, X., Lin, Z., Liu, J., Wang, Y., Feng, J., Yang, M.H., Liew, J.H.: QK-Edit: Revisiting Attention-based Injection in MM-DiT for Image and Video Editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19043–19053 (2025)
64. Song, Y., Huang, S., Yao, C., Ci, H., Ye, X., Liu, J., Zhang, Y., Shou, M.Z.: Processpainter: Learning to draw from sequence data. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–10 (2024)
65. Song, Y., Liu, C., Jiang, Y., Shou, M.Z.: Streamingeffect: Real-time human-centric video effect generation. arXiv preprint arXiv:2605.17019 (2026)
66. Song, Y., Yao, W., Wang, H., Shou, M.Z.: Vista: Triplet-supervised video style transfer with diffusion transformers. arXiv preprint arXiv:2605.17312 (2026)
67. Sun, W., Tu, R.C., Liao, J., Tao, D.: Diffusion Model-Based Video Editing: A Survey. arXiv preprint arXiv:2407.07111 (2024)
68. Tu, S., Dai, Q., Cheng, Z.Q., Hu, H., Han, X., Wu, Z., Jiang, Y.G.: MotionEditor: Editing Video Motion via Content-Aware Diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7882–7891 (2024)
69. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint arXiv:2503.20314 (2025)

70. Wang, J., Ma, Y., Guo, J., Xiao, Y., Huang, G., Li, X.: Cove: Unleashing the diffusion feature correspondence for consistent video editing. Advances in Neural Information Processing Systems **37**, 96541–96565 (2024)
71. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. arXiv preprint arXiv:2411.04746 (2024)
72. Wang, W., Jiang, Y., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-Shot Video Editing Using Off-the-Shelf Image Diffusion Models. arXiv preprint arXiv:2303.17599 (2023)
73. 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)
74. Wu, Y., Cheng, H., He, Z., Liu, S.: Vibe: Ultra-high-resolution video synthesis born from pure images. arXiv preprint arXiv:2603.23326 (2026)
75. Wu, Y., Song, J., Tan, Z., He, Z., Liu, S.: Freeswim: Revisiting sliding-window attention mechanisms for training-free ultra-high-resolution video generation. arXiv preprint arXiv:2511.14712 (2025)
76. Xu, X., Li, Y., You, S., Bao, B.K.: Smrabooth: Subject and motion representation alignment for customized video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16130–16141 (2026)
77. 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)
78. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: UNIC: Unified In-Context Video Editing. In: Proceedings of the International Conference on Learning Representations (2026)
79. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation. arXiv preprint arXiv:2303.12346 (2023)
80. Yu, M., Hu, W., Xing, J., Shan, Y.: TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 100–111 (2025)
81. 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)
82. Zheng, S., Feng, L., Wang, X., Zhou, Q., Cai, P., Zou, C., Liu, J., Lin, Y., Chen, J., Ma, Y., et al.: Forecast then calibrate: Feature caching as ode for efficient diffusion transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13449–13457 (2026)
83. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-Sora: Democratizing Efficient Video Production for All. arXiv preprint arXiv:2412.20404 (2024)