CLEVAM-DM: Consistent Local Edits in Videos via Attention Manipulation in Diffusion Models

Hanzuo Liu
IIIS, Tsinghua University
lhz24@mails.tsinghua.edu.cn

Kehan Liu
IIIS, Tsinghua University
liukh24@mails.tsinghua.edu.cn

Zehua Wang
IIIS, Tsinghua University
wangzehu24@mails.tsinghua.edu.cn

Abstract

We introduce CLEVAM-DM, a training-free framework for temporally consistent local video editing via attention manipulation in diffusion models. Our pipeline combines keyframe-based inpainting, full attention sharing, and perceptually guided frame interpolation to achieve high-quality, coherent edits while preserving motion and background. CLEVAM-DM operates efficiently on consumer hardware, requires no additional training, and is compatible with existing diffusion models. Experiments show remarkable consistency and fidelity across diverse videos. Our code is available at https://github.com/patrickwzh/CLEVAM-DM.

1 Introduction

Video editing with generative AI presents distinct technical challenges beyond image editing, primarily due to the critical requirement for temporal coherence across frames. While diffusion models have demonstrated exceptional capabilities in image generation and manipulation [1], their direct application to video editing frequently introduces temporal inconsistencies or undesirable global alterations [2]. Current methodologies can be broadly categorized into two paradigms: (1) direct video processing approaches that typically demand substantial computational resources, and (2) keyframe-based methods that often fail to maintain original motion characteristics when performing localized object modifications.

Recent developments in controlled diffusion techniques, particularly attention manipulation [3] and style alignment mechanisms [4], offer promising solutions to these challenges. However, these advancements have been predominantly limited to static image applications, lacking direct applicability to video processing. Our research addresses this critical gap by introducing a novel framework that achieves consistent local video edits while preserving both motion dynamics and background elements.

The principal technical challenges in consistent local video editing include:

- **Temporal Coherence Maintenance**: Ensuring edited objects exhibit consistent visual attributes throughout the temporal sequence
- **Motion Fidelity Preservation**: Maintaining original motion trajectories while modifying object appearances
- **Precision in Text-Guided Generation**: Achieving accurate alignment between textual prompts and visual modifications

Preprint. Under review.

- **Computational Tractability**: Delivering high-quality results within practical computational constraints

This paper presents CLEVAM-DM, an innovative training-free pipeline that systematically addresses these challenges through three key technical contributions:

- **Model-Agnostic Architecture**: Our framework demonstrates compatibility with diverse image-based diffusion models, facilitating seamless integration of emerging advancements
- **Zero-Shot Adaptation**: Unlike conventional video editing approaches, our method operates without requiring additional model training
- **Computational Efficiency**: The complete pipeline achieves state-of-the-art performance on a single consumer-grade GPU with 24GB memory

Comprehensive experimental evaluations demonstrate that CLEVAM-DM consistently produces high-fidelity local edits across diverse video content while maintaining superior temporal consistency and environmental preservation compared to existing approaches.

2 Related Work

2.1 Diffusion Models for Video Editing

Recent advances in diffusion models have revolutionized video editing by enabling high-fidelity generation and manipulation. The field has evolved along three primary directions: temporal adaptation, structure conditioning, and training-free methods. Temporal adaptation approaches like Fairy [5] employ 3D spatio-temporal attention for parallelized keyframe editing, while VidToMe [6] improves efficiency through token merging. Structure conditioning methods such as FlowVid [7] and MoCA [8] explicitly integrate optical flow to maintain motion consistency, though they require computationally expensive flow estimation during inference. Training-free paradigms including Text2Video-Zero [9] and Vid2Vid-Zero [10] achieve zero-shot editing through latent warping or attention injection. Our work advances these directions by combining DDIM inversion with cross-frame attention sharing, eliminating the need for explicit flow estimation of processed video while maintaining superior local consistency.

2.2 Attention Mechanisms in Video Generation

Attention manipulation has emerged as a powerful tool for temporal consistency in video generation. TokenFlow [11] and FLATTEN [12] utilize optical flow or feature similarity to guide attention across frames, requiring explicit motion modeling. UniEdit [13] and AnyV2V [3] employ multi-branch architectures to disentangle appearance and motion, but suffer from increased memory requirements. While recent work like Shape-guided diffusion [3] applies shared attention only on the first layer of the U-Net, we demonstrate that full attention sharing across all network layers yields superior performance.

2.3 Local Editing Techniques

Local video editing faces unique challenges in preserving both content consistency and motion dynamics. Traditional approaches rely on frame-by-frame processing with post-hoc temporal alignment [2] or dedicated video diffusion architectures [14, 15]. Alternative methods like LNA [16] and CoDeF [17] employ canonical 2D atlases for video editing, introducing additional complexity in atlas learning and mapping. PYoCo [18] demonstrates the effectiveness of latent clustering for video frames, suggesting the potential of proper latent initialization. Our method builds upon these insights while avoiding their limitations through direct feature manipulation via masked diffusion and DDIM-inverted latent initialization.

2.4 Frame Interpolation and Motion Preservation

Frame interpolation techniques have progressed significantly [19] from traditional optical flow methods [20] to modern deep learning approaches. While flow-free methods like DAIN [21] bypass

explicit flow computation, they require extensive training data. Hybrid approaches such as PerVFI [22] combine flow estimation with perceptual losses for robust performance. Time-aware architectures including AdaCoF [23] address motion ambiguity through adaptive flow blending. Rerender-A-Video [24] introduces dual-flow fusion but requires per-frame alpha blending. Our proposed system combines PerVFI with original flow priors and motion-consistent blending, achieving artifact-free interpolation without manual mask tuning.

3 Methodology

Our framework, CLEVAM-DM, operates through three sequential stages: keyframe preparation, consistent local editing, and video frame interpolation. Figure 1 provides an overview of the complete pipeline.

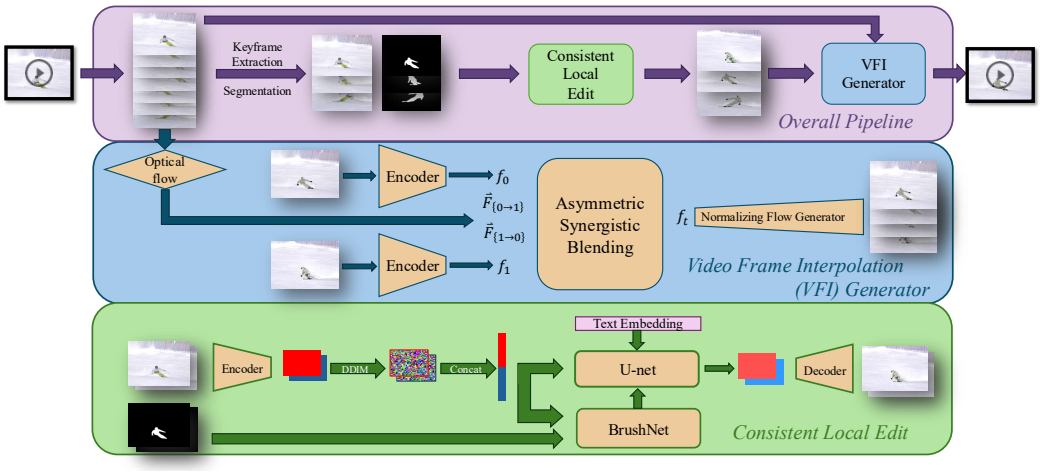


Figure 1: Overview of the CLEVAM-DM workflow. **Top:** The overall pipeline extracts keyframes, applies consistent local edits, and synthesizes intermediate frames. **Middle:** Video frame interpolation module utilizes the optical flow of the original video, and generates smooth transitions between edited keyframes. **Bottom:** Consistent local editing is achieved via inpainting and attention sharing in diffusion models.

3.1 Keyframe Extraction and Mask Inference

The first stage of our pipeline focuses on identifying representative keyframes from the source video and generating accurate masks for the target objects to be edited. Keyframe selection is performed at regular intervals, balancing the need to capture essential motion dynamics with computational efficiency. The interval is typically chosen based on a user-specified sample rate.

For each selected keyframe, we employ Language Segment Anything [25], a text-guided segmentation model, to automatically generate precise object masks. This approach leverages user-provided textual prompts to specify the target object, enabling flexible and intuitive mask generation without manual annotation. By incorporating language-driven segmentation, our method reduces ambiguity in object selection and ensures that the masks accurately correspond to the intended editing regions.

The resulting set of keyframes, each paired with a high-quality segmentation mask, forms the foundation for subsequent editing operations. Figure **??** illustrates the segmentation and mask inference process, highlighting how textual prompts such as “in the front” can help disambiguate object selection in complex scenes.

3.2 Local Edits through Inpainting

For local editing of keyframes, we leverage BrushNet [26], an inpainting model built on pretrained diffusion backbones. BrushNet employs a dual-branch architecture: the feature extraction branch

takes concatenated noisy latents, masked image latents, and a downsampled mask to produce hierarchical features, which are then inserted layer-by-layer into the frozen diffusion U-Net through zero convolution blocks. This design enables precise control over generation and allows adjusting the preservation scale of unmasked regions via the feature insertion weight. A blending operation using a blurred version of the mask in pixel space ensures seamless transitions between edited and preserved areas.

We adopt DDIM inversion [27] to initialize latents that retain the structural and motion cues of the original frames. This combination of BrushNet's hierarchical feature insertion and DDIM-inverted latents provides flexible, high-fidelity edits—foreground objects or background elements can be targeted simply by inverting the segmentation mask—while maintaining temporal consistency.

Figure 2 illustrates our modified BrushNet architecture and the integration of DDIM inversion for local video editing.

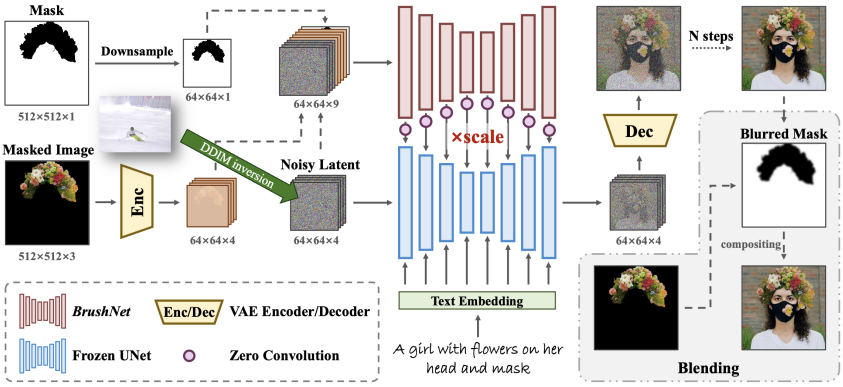


Figure 2: BrushNet model architecture for local video editing with DDIM-inverted latents.

3.3 Temporal Consistency Enhancement

To ensure consistent appearance across edited keyframes, we implement a full attention sharing mechanism [28] across all attention layers in the BrushNet model. While Hertz et al. [28] originally restricted attention sharing to a single reference image to avoid content leakage between unrelated images, our scenario differs: all frames depict the same object under varying poses or contexts. Thus, we enable full attention sharing across all keyframes, promoting maximal alignment of appearance features throughout the sequence.

Concretely, during the diffusion process, we share attention keys and values across all frames being processed in parallel. For each frame $\mathcal{I}_i$ with deep features ϕ_i, the queries Q_i, keys K_i, and values V_i are computed as usual. However, instead of computing attention using only the keys and values from the current frame, we concatenate the keys and values from all n frames in the batch. The attention update for ϕ_i is then given by:

$$\text{Attention}(Q_i, K_{1\ldots n}, V_{1\ldots n}),$$

where $K_{1\ldots n} = \begin{bmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{bmatrix}$ and $V_{1\ldots n} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}$. This formulation encourages the model to generate consistent features and appearance for the edited regions across all frames.

By leveraging this full attention sharing strategy, our method enforces strong temporal consistency in local edits, ensuring that object appearance remains coherent throughout the video sequence. This approach, illustrated in Figure 3, enables consistent local editing without requiring explicit video diffusion models or additional temporal regularization losses. The temporal consistency is shown in Figure 3 and in the ablation study (Figure 6).

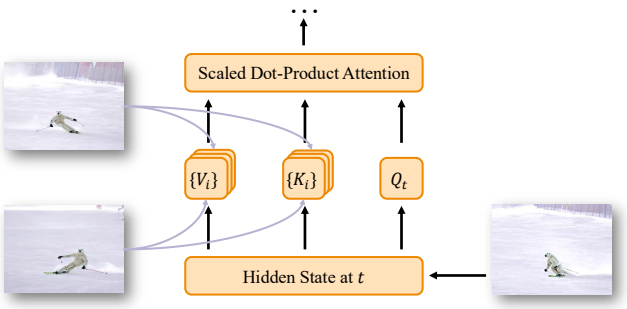


Figure 3: Temporal consistency enhancement through full attention sharing in the BrushNet model.

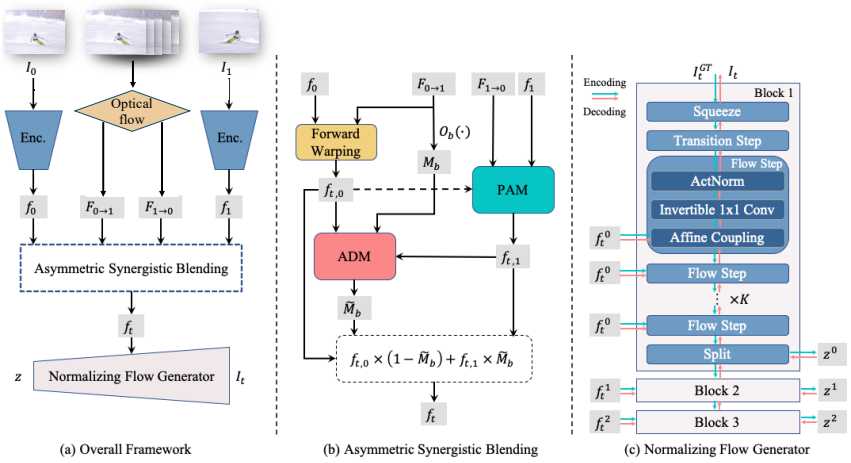
3.4 Video Frame Interpolation

The final stage of our pipeline focuses on generating intermediate frames between the edited keyframes to reconstruct a temporally coherent video sequence. To achieve this, we first estimate optical flow between pairs of edited keyframes, which serves as a guide for the interpolation process. Crucially, we also incorporate optical flow information from the original video, as it provides more accurate and reliable motion cues than can be inferred from the edited frames alone. This dual-flow strategy enables the interpolation module to better preserve the original motion dynamics and handle challenging scenarios such as rapid movements or occlusions.

For the interpolation itself, we utilize PerVFI [22], a perception-oriented video frame interpolation model that has demonstrated state-of-the-art performance in generating visually plausible intermediate frames. PerVFI leverages deep neural networks to synthesize high-quality frames that maintain both structural integrity and perceptual realism.

To further enhance the fidelity of the interpolated frames and minimize artifacts, we introduce a selective region preservation mechanism. Specifically, for each interpolated frame, we identify unmasked regions—areas that were not subject to editing—and directly copy these regions from the corresponding original frames. This hybrid approach ensures that unedited content remains consistent with the source video, significantly reducing ghosting and blending artifacts that can arise during interpolation. The effect of this mechanism is shown in ablation studies 8.

By integrating optical flow from both the original and edited videos, employing a perceptually optimized interpolation model, and selectively preserving unedited regions, our method achieves high-quality, temporally consistent video sequences. Figure 4 illustrates the overall frame interpolation process and highlights the effectiveness of our approach in maintaining both motion continuity and visual fidelity.



(a) Overall Framework (b) Asymmetric Synergistic Blending (c) Normalizing Flow Generator

Figure 4: Frame interpolation process using optical flow from the original video and PerVFI.

4 Experiments

4.1 Experimental Setup

We evaluate CLEVAM-DM on a diverse set of videos encompassing various subjects, motion patterns, and backgrounds. All experiments are conducted on a single NVIDIA GeForce RTX 3090 or RTX 4090 GPU. The diffusion backbone is Stable Diffusion v1.5 as integrated in BrushNet. Text-guided segmentation is performed using LangSAM [25], and frame interpolation is carried out with PerVFI [22].

4.2 Qualitative Results

Figure 5 presents qualitative results of CLEVAM-DM on representative video sequences. The method achieves temporally consistent local edits, preserving both object appearance and motion dynamics. Edited regions exhibit high visual fidelity, while unedited areas remain unchanged, demonstrating the effectiveness of our selective region preservation strategy.

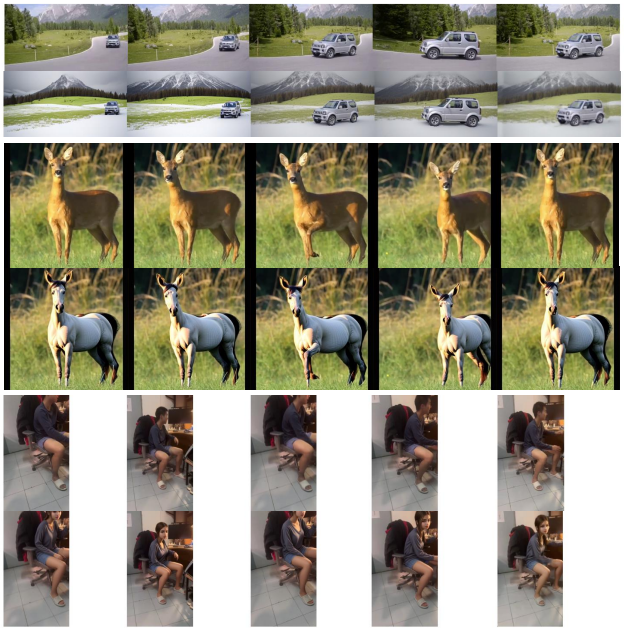


Figure 5: Qualitative results of CLEVAM-DM. **Top**: original video frames. **Bottom**: processed frames. The first four processed frames are keyframes; the last column shows interpolated frames.

4.3 Computation Time Analysis

Table 1 presents the typical computation time required for each component of our pipeline when processing a 5-second 720p video. CLEVAM-DM achieves efficient processing, with per-frame inpainting and interpolation times suitable for practical use. Compared to end-to-end video diffusion models, our approach offers significant computational advantages.

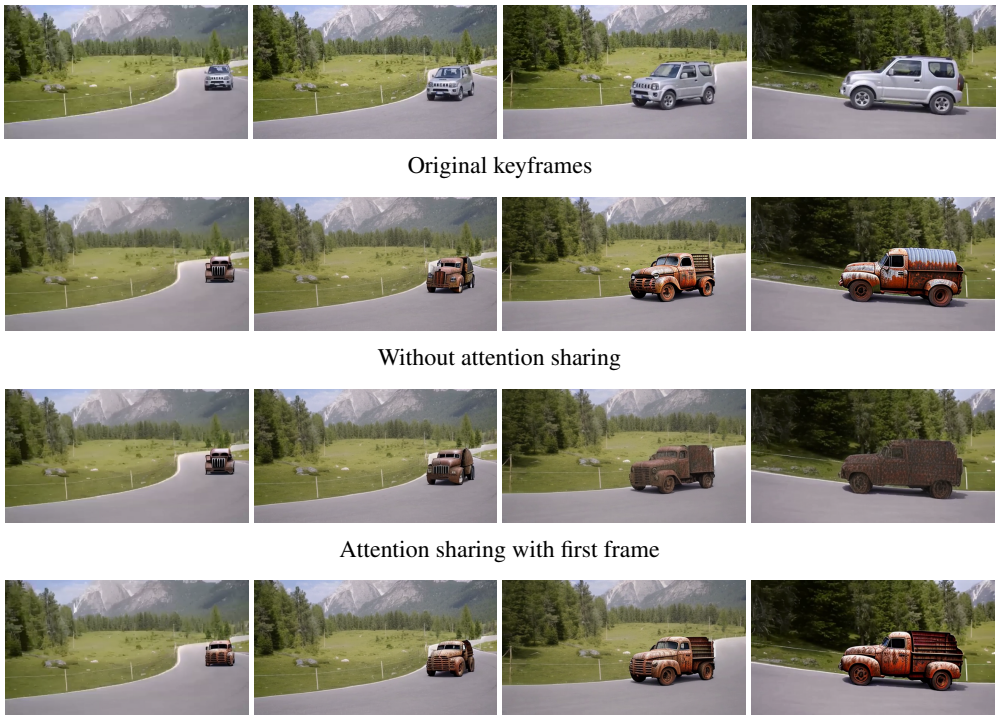
Table 1: Computation time analysis of CLEVAM-DM components.

Component	Time (seconds)
Keyframe extraction	2
Text-guided segmentation	80
Consistent local edit	150
Frame interpolation	60

4.4 Ablation Studies

We perform ablation studies to evaluate the impact of each component in our pipeline. All experiments are conducted using the prompt "car" (source) to "a rusty truck" (target), with only the relevant factor varied in each ablation while keeping all other settings fixed.

Effect of Attention Sharing. We investigate the impact of attention sharing by comparing three approaches: full attention sharing across all frames, independent processing of each frame, and attention sharing limited to the first frame using AdaIN [28]. As demonstrated in Figure 6, full attention sharing markedly enhances temporal consistency and reduces appearance variations across frames.



Original keyframes

Without attention sharing

Attention sharing with first frame

Full attention sharing

Figure 6: Comparison of edited frames with and without attention sharing. Attention sharing ensures consistent appearance across frames.

Impact of DDIM Inversion. We evaluate the effect of DDIM inversion on structural preservation. Figure 7 shows that DDIM inversion improves alignment between original and edited frames, maintaining pose and motion.

(a) Original (b) Without Inversion (c) With Inversion

Figure 7: Comparison of edited frames with and without DDIM inversion. DDIM inversion preserves the original structural information and improves alignment.

Effect of Selective Region Preservation. Selectively preserving unmasked regions from the original frames during interpolation reduces ghosting artifacts and maintains fidelity in unedited areas (Figure 8).

Without region preservation

With region preservation

Figure 8: Comparison of interpolated frames with and without selective preservation of unmasked regions. Selective preservation reduces ghosting artifacts, which appear as blurry color patches around the car—mostly with the color of the rusty truck.

5 Conclusion

We have introduced CLEVAM-DM, a novel framework for achieving temporally consistent local edits in videos by leveraging attention manipulation within diffusion models. Our approach integrates advanced inpainting techniques, full attention sharing mechanisms, and perceptually guided frame interpolation to enable high-quality, temporally coherent local edits while preserving both motion dynamics and background integrity.

CLEVAM-DM distinguishes itself from prior work by operating efficiently on consumer-grade hardware, requiring no additional training, and maintaining compatibility with a wide range of image-based diffusion models. Experimental results demonstrate that our method produces visually consistent and high-fidelity edits across diverse video content, supporting both foreground and background modifications.

The proposed pipeline offers several advantages: it delivers temporally consistent local edits, preserves original motion patterns, and is adaptable to various editing scenarios without the need for specialized training or hardware. These characteristics make CLEVAM-DM a practical and effective solution for video editing tasks in both research and real-world applications.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[2] Yuanzhi Wang, Yong Li, Xiaoya Zhang, Xin Liu, Anbo Dai, Antoni B. Chan, and Zhen Cui. Edit temporal-consistent videos with image diffusion model, 2023.

[3] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention, 2024.

[4] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention, 2024.

[5] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8261–8270, June 2024.

[6] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7486–7495, June 2024.

[7] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8207–8216, June 2024.

[8] Wilson Yan, Andrew Brown, Pieter Abbeel, Rohit Girdhar, and Samaneh Azadi. Motion-conditioned image animation for video editing, 2023.

[9] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, October 2023.

[10] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models, 2024.

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *ICLR*, 2024.

[12] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zheng Zhong, Weijia Liu, Pengfei Li, Jingren Yang, and Jiaya Lu. Flatten: Optical flow-guided attention for consistent text-to-video editing. *NeurIPS*, 2023.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, et al. Prompt-to-prompt image editing with cross-attention control. *ICLR*, 2023.

[14] Uriel Singer, Adam Polyak, Thomas Hayes, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023.

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, et al. Video diffusion models. *NeurIPS*, 2022.

[16] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. 40(6), 2021.

[17] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8089–8099, June 2024.

[18] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22930–22941, October 2023.

[19] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. Diffusion model-based video editing: A survey. *arXiv preprint arXiv:2407.07111*, 2024.

[20] Huaizu Jiang, Deqing Sun, Varun Jampani, et al. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *CVPR*, 2018.

[21] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Ma, Yingyu Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, 2019.

[22] Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. Perception-oriented video frame interpolation via asymmetric blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2753–2762, 2024.

[23] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *CVPR*, 2020.

[24] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender-a-video: Improving video consistency via diffusion-based full attention sharing. *NeurIPS*, 2023.

[25] Luca Medeiros. Language segment anything. `https://github.com/luca-medeiros/lang-segment-anything`, 2023. Accessed: 2025-05-29.

[26] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[28] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Stylealigned: Learning to share attention for consistent image generation. *ACM Transactions on Graphics (TOG)*, 43(2):1–12, 2024.