LLM-DRIVEN VIDEO INPAINTING WITH EXPLICIT MASK GUIDANCE AND WARP-RELATION CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Video inpainting is a fundamental task with wide applications in film post-production and object removal. Existing text-guided image and video editing methods typically rely on implicit conditioning by injecting text embeddings into the generation process, which lacks explicit intermediate representations and makes it difficult to precisely align the semantic space with the pixel space. To address this limitation, we propose an LLM-guided video inpainting framework that leverages a Multi-Modal large language model to generate explicit masks, followed by a mask smoothing and enhancement module for post-processing, and a video inpainting backbone for final restoration. Furthermore, We propose a Warp-Relation Consistency Mechanism that explicitly enforces temporal alignment between frames via flow-guided warping and relation-aware constraints. Extensive experiments demonstrate that our approach not only achieves state-of-the-art PSNR and SSIM, but also effectively reduces mask boundary artifacts and improves temporal consistency compared to existing methods.We will publicly release the code and pretrained models to facilitate reproducible research.

1 INTRODUCTION

Video inpainting is a traditional video restoration task, whose aim is to restore the video by filling in the surrounding and previous relevant content of the damaged area Bertalmio et al. (2001). Its application scenarios are very extensive, such as post-production of films and television Lin et al. (2024), removal of objects Li et al. (2022), video restorationChen et al. (2023), creative editing Guo et al. (2023), autonomous driving Buburuzan et al. (2025), etc. With the rise of short videos, their audience base is expanding and the system is developing rapidly. Generally speaking, they can be roughly divided into two categories: unrestricted video inpainting and restricted video inpainting.

Most traditional methods Liu et al. (2021) Ren et al. (2022) Wu et al. (2024b) Zhang et al. (2024) fall into the category of unrestricted conditions and usually rely only on mask sequences. Under the guidance of optical flow, they fill in the missing areas through the transmission and generation mechanism of spatio-temporal information. This method is relatively dependent on the existing content around the missing area or in the previous and subsequent frames, and requires precise masking positioning, which limits the application range of users.

In contrast, the video inpainting method with constraints not only relies on the existing pixel information of the input video when generating the content of the missing area, but also introduces additional prior information or external guidance as constraints, thereby enhancing the semantic consistency and visual quality of the generated results. For example, Cho et al. (2025)proposed taking the first frame as the key frame, first performing high-quality restoration on the key frame, and then using optical flow information to propagate the restoration result of the key frame to subsequent frames to achieve cross-frame time-consistent inpainting. The advantage of this method lies in its ability to significantly reduce the drift and flicker phenomena generated in subsequent frames. SAVIT Lee et al. (2023)adopts the semantic-aware Dynamic Experts mechanism to perform hyperclassification of intra-frame objects before inpainting and assign Semantic labels to different objects. Subsequently, these semantic labels are utilized to guide the content generation of the missing areas, thereby ensuring the semantic consistency between the generated areas and the original scene. Such methods typically rely on precise mask segmentation as a guiding condition, enabling the model to clearly know which regions need to be repaired and which remain as they are, thereby better controlling the generation quality and semantic fidelity. Methods such as Zi et al. (2025) and Brooks et al. (2023) take text input as a constraint condition and guide the generation process in the semantic space through CLIP embedding to achieve semantic control of video content. However, most of these methods focus on video editing rather than specialized video inpainting. Their goal is to generate or modify existing content rather than simply repair damaged areas. To overcome these limitations, some latest studies Wu et al. (2024a) have begun to take Multi-Modal large language models into consideration by parsing the natural language prompts input by users. And inject the generated semantic information into the Cross Attention module of the diffusion model to guide video generation. Although this approach can achieve semantic-driven content modification, since language information mainly functions in the semantic space and lacks precise alignment with the pixel space, it is difficult to perform post-processing operations such as boundary optimization or morphological smoothing. Furthermore, these methods typically do not explicitly constrain cross-frame consistency, which leads to the restoration results being prone to flickering, ghosting or instability in the temporal dimension, thereby limiting their application in high-quality video restoration and inpainting tasks.

In this work, we propose an LLM-guided framework for video inpainting, which fully utilizes the semantic understanding capabilities of large language models to transform users' text instructions into explicit mask constraints. Specifically, we use MLLM to automatically generate the initial mask of the target area and perform post-processing optimization on the mask through Gaussian blur and binarization operations, thereby enhancing the repair effect simultaneously at the semantic level and the pixel level. Unlike the existing methods that directly inject language information into the generative network, our approach can explicitly control the area that needs to be restored, providing clear spatial guidance for the generative model during inpainting, thereby enhancing the accuracy and controllability of the restoration.

Furthermore, we proposed the Warp-Relation Consistency (WRC) mechanism to enhance the consistency of videos across frames. This mechanism combines optical flow-guided inter-frame distortion and relational awareness constraints to explicitly constrain the temporal alignment between video frames, thereby effectively alleviating common problems in video restoration such as flicker, glint, or discontinuous object morphology. Through the WRC mechanism, the model can not only generate high-quality single-frame restoration results, but also ensure the natural continuity of the restoration object in the temporal dimension, making the video visually smoother and more stable.

In summary, this work achieves precise control and high-quality generation of video inpainting by integrating the semantic understanding capabilities of MLLM and explicit cross-frame consistency constraints, significantly enhancing semantic alignment and temporal stability, and providing an effective new method for text-driven video restoration.

2 Related Work

2.1 Video inpainting

Traditional video inpainting methods mainly focus on restoring damaged areas by utilizing limited spatial and temporal information. For this goal, multiple technical routes have been developed. Some methods Zeng et al. (2020) Bian et al. (2025), without the aid of optical flow, perform feature encoding and generation on the missing regions through spatio-temporal transformation, which can maintain inter-frame consistency to a certain extent. Clark et al. (2017) adopts CNN to simulate the FlowNet structure and simultaneously uses LSTM to encode time series information to predict and fill the missing areas, thereby achieving content completion across frames.

However, with the increasing complexity of video scenes, most mainstream methods Xu et al. (2019)Zhang et al. (2022)Gao et al. (2020)Henschel et al. (2025)Wang et al. (2025) rely on optical flow to capture the precise motion of objects in the previous and subsequent frames, thereby enhancing temporal consistency and generation effect. More crucially, almost all of these methods require precise masks to identify the areas that need to be repaired. Neither optical flow-assisted nor non-optical flow-assisted methods can bypass this constraint. At the video scale, the annotation and processing of masks often consume a large amount of manual and computing resources, which not

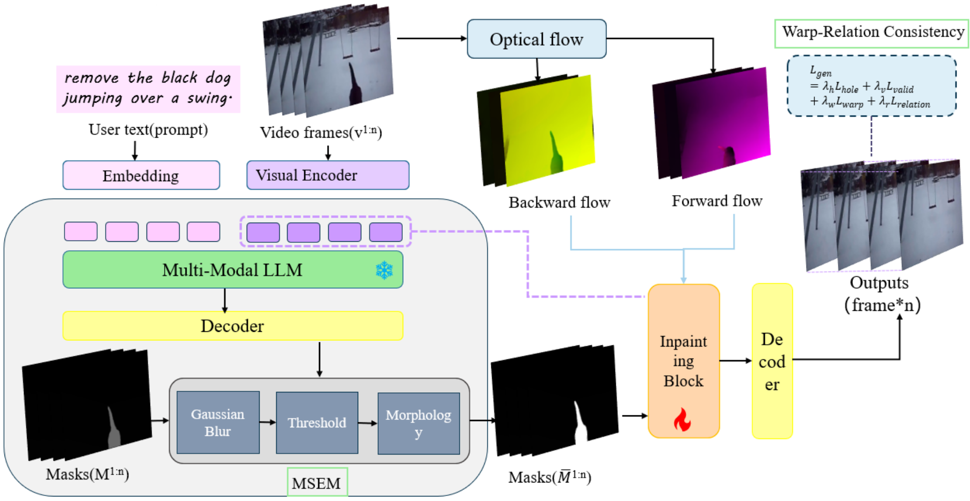


Figure 1: Overview pipeline of our MaRC

only increases the usage cost but also limits the scalability and universality of the method in actual scenarios.

2.2 Text-guided video editing and inpainting

Text demonstrates great application potential and research value in the field of video generation and editing due to its convenient accessibility and low cost. Zi et al. (2025) integrates a Stable Diffusion-based Image Inpainting model with a Personalized Editing model, and incorporates motion modules at each intermediate layer following the principle of task vectors to enforce temporal consistency. Khachatryan et al. (2023) introduces a warp function that propagates latent features across frames to simulate camera and scene transformations, thereby extending image editing models to the video domain.

Models such as Singer et al. (2022), Mei & Patel (2023), and Wu et al. (2023) focus on text-guided video generation, where language information guides the content creation. These approaches emphasize generation rather than restoration, and thus may produce artifacts or generate undesired new content when applied to video inpainting tasks. In contrast, Wu et al. (2024a) specifically targets video inpainting, and for the first time leverages the powerful text understanding capabilities of Multi-Modal large language models to guide video restoration.

The above methods, whether CLIP-guided Radford et al. (2021) or MLLM-guided, typically inject semantic embeddings directly into the cross-attention module for computation. However, this strategy suffers from insufficient alignment between the semantic space and pixel space, making precise boundary control challenging. Moreover, these methods generally lack effective temporal constraints, often leading to flickering or discontinuities in the generated videos.

3 Methods

3.1 Preliminary

3.1.1 LISA for Image Inpainting

Large language models have been successfully applied to image inpainting, achieving strong semantic understanding. We build on the LISA model Lai et al. (2024), which introduces a special ¡seg¿ token to indicate editable regions. LISA freezes the vision backbone and uses an LLM to parse user prompts into segmentation instructions, which are then executed by the Segment Anything Model Kirillov et al. (2023) to produce accurate masks for image editing.

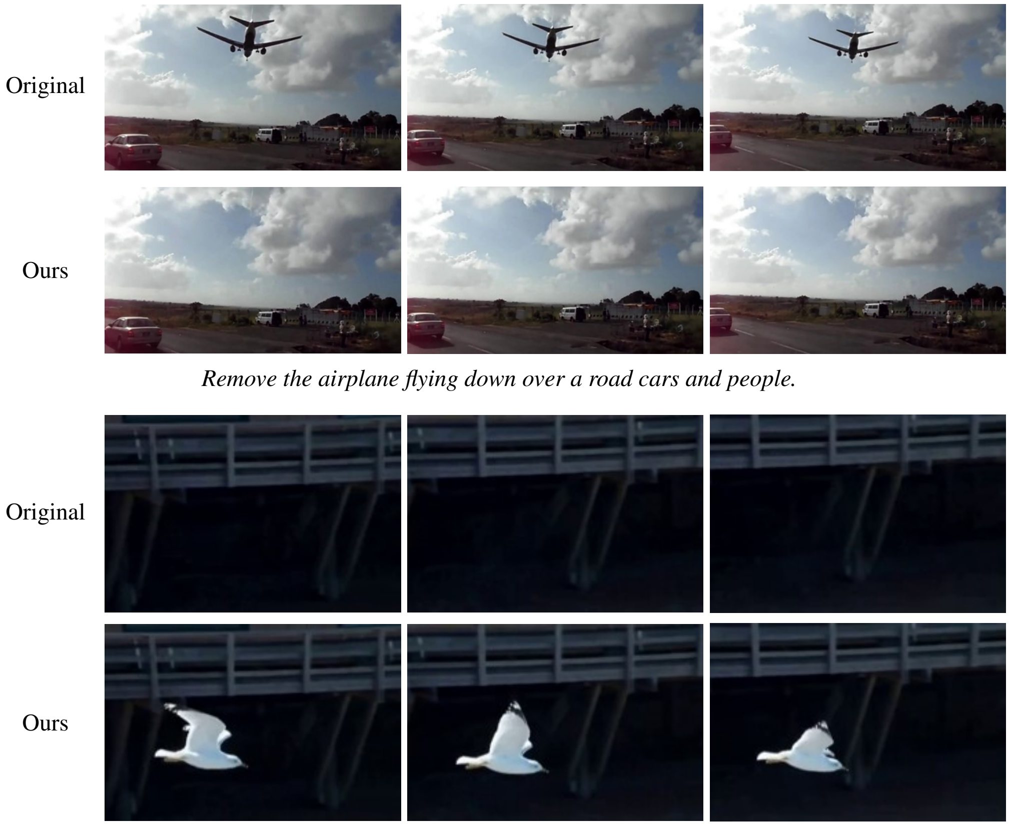
3.1.2 ProPainter for Video Inpainting

To perform inpainting within video sequences, we adopt ProPainter Zhou et al. (2023) as the inpainting block. ProPainter is a state-of-the-art video inpainting framework that leverages feature propagation and temporal alignment to fill in missing regions across frames with high consistency. It introduces bidirectional propagation to exploit both past and future contexts, and employs a feature alignment module to mitigate motion mismatch between adjacent frames. This design enables the model to maintain temporal coherence while producing visually plausible content in occluded regions.

By integrating ProPainter as our inpainting backbone, we ensure that once the masks are generated , the missing regions can be restored with both spatial fidelity and temporal consistency.

3.1.3 Overview

To address the misalignment between semantic and pixel spaces in text-guided video inpainting and reduce manual effort, we leverage LISA as a preliminary module to generate masks from prompts. These masks are refined in the MSEM module to improve boundaries and fill internal holes, then fed into the inpainting module for content restoration. Sequential application of image-based LISA can introduce temporal inconsistencies, so we further propose Warp-Relation Consistency to enforce coherence across frames. This pipeline effectively combines language-guided segmentation, mask refinement, and temporally consistent inpainting for high-quality video restoration.The complete framework diagram is shown in Fig 1.



Remove the airplane flying down over a road cars and people.

Remove the white seagull flying in the city.

Figure 2: Qualitative results of MaRC. Each pair of rows corresponds to a video prompt, showing Original and Ours frames.

Given a video sequence

$$V = \{v_1, v_2, \dots, v_T\}, \quad v_t \in \mathbb{R}^{H \times W \times 3}, \tag{1}$$

and a textual prompt $p \in \mathcal{P}$, our goal is to automatically generate frame-wise masks

$$M = \{m_1, m_2, \ldots, m_T\}, \quad m_t \in [0,1]^{H \times W}, \tag{2}$$

indicating regions to be restored.

Initial Mask Generation Each frame v_t is encoded by a visual encoder E_v to obtain features:The textual prompt p is tokenized and wrapped with a special `<seg>` token, then embedded by a language model E_l to produce semantic guidance:

These visual and textual features are fed into a Multi-Modal large language model combined with the Segment Anything Model to generate the initial mask:

$$m_t^{\text{raw}} = \text{SAM}(\text{MLLM}(f_t, h_p)). \tag{3}$$

The raw mask m_t^{raw} roughly indicates the regions described by the prompt but is often coarse, jagged, and may contain internal holes due to imperfect semantic-to-pixel alignment.

Mask Refinement To produce masks suitable for inpainting, we apply a three-step refinement procedure inspired by morphological operations in our MSEM model:

Gaussian Blur: smooths jagged edges and high-frequency noise:

$$\tilde{m}_t^{(1)} = G_\sigma * m_t^{\text{raw}}, \tag{4}$$

where G_σ is a Gaussian kernel and $*$ denotes convolution.

Thresholding: converts the blurred mask into a binary mask:

$$\tilde{m}_t^{(2)}(i,j) = \begin{cases} 1, & \tilde{m}_t^{(1)}(i,j) > \tau \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where τ is a predefined threshold to remove low-confidence regions.

Morphological Operations: removes small holes and connects fragmented regions:

$$m_t^{\text{refined}} = \text{Morph}(\tilde{m}_t^{(2)}), \tag{6}$$

where $\text{Morph}(\cdot)$ includes operations such as closing (dilation followed by erosion) and optional opening to eliminate isolated noise.

The final refined mask sequence is

$$M^{\text{refined}} = \{m_1^{\text{refined}}, m_2^{\text{refined}}, \ldots, m_T^{\text{refined}}\}, \tag{7}$$

which is then fed into the inpainting module. This pipeline combines semantic guidance from the LLM, pixel-level segmentation via SAM, and morphological post-processing to produce accurate and spatially coherent masks suitable for temporally consistent video inpainting.

3.2 Warp-Relation Consistency

Extending image inpainting modelsYang et al. (2024)Wang et al. (2018) to video often introduces temporal inconsistency due to moving objects and camera motion. Existing approaches enforce temporal consistency by warping features using optical flow and computing feature differences between adjacent frames. While effective in reducing abrupt changes, these losses mix natural motion with reconstruction errors, which can lead to suboptimal results in moving regions.

Inspired by the relation-based temporal consistency proposed in Dai et al. (2022), we adapt this idea for video inpainting by introducing a **mask-restricted Warp-Relation Consistency** loss. Specifically, we first warp features from adjacent frames to the current frame and compute the differences between predicted and ground-truth warped features. By restricting this computation to the inpainting mask or other regions of interest, the loss focuses on areas where reconstruction errors are expected, avoiding inappropriate penalties in moving background regions and preserving natural motion.

Formally, the loss is defined as:

$$L_{\text{relation}} = \sum_t \left\| (F_t - W(F_{t-1})) - (F_t^{GT} - W(F_{t-1}^{GT})) \right\|_\Omega, \tag{8}$$

where F_t denotes predicted features, F_t^{GT} denotes ground-truth features, $W(\cdot)$ is the optical flow warp operation, and Ω indicates the mask or regions of interest.

In addition, we include a standard **warp loss**:

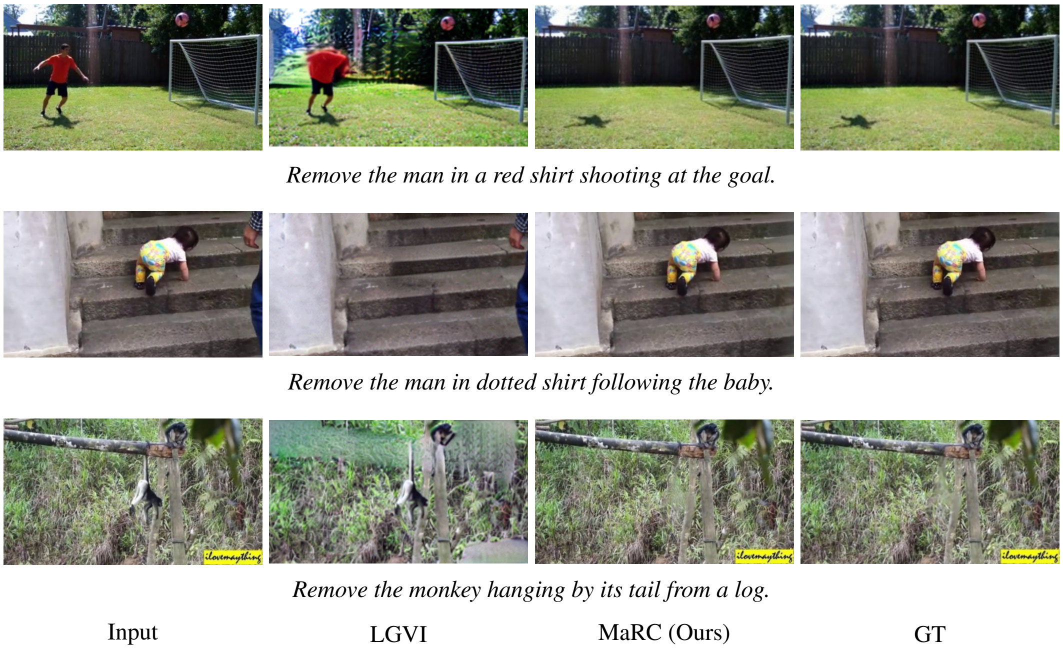
$$L_{\text{warp}} = \sum_t \left\| F_t - W(F_{t-1}) \right\|_\Omega. \tag{9}$$

While L_{warp} reduces abrupt changes and flickering by directly enforcing feature consistency between adjacent frames, it does not distinguish reconstruction errors from natural motion. By combining L_{warp} with our mask-restricted Warp-Relation Consistency loss, we achieve both smooth temporal transitions and accurate reconstruction in masked regions.

The overall training objective is:

$$L_{\text{gen}} = \lambda_h L_{\text{hole}} + \lambda_v L_{\text{valid}} + \lambda_w L_{\text{warp}} + \lambda_r L_{\text{relation}}, \tag{10}$$

where L_{hole} and L_{valid} are conventional pixel-wise losses in masked and unmasked regions, respectively. λ_h, λ_v, λ_w, and λ_r are weighting factors.



Remove the man in a red shirt shooting at the goal.

Remove the man in dotted shirt following the baby.

Remove the monkey hanging by its tail from a log.

Figure 3: Quantitative Comparison of text-guided video inpainting. Each row shows one frame, with the prompt on the left. Column labels at the bottom.

4 Experiment

4.1 Implementation Details

4.1.1 Dataset

For all experiments, we use the ROVI Wu et al. (2024a) dataset, which is specifically designed for text-guided video inpainting tasks. The dataset consists of 5,650 video clips and 9,091 inpainting samples. Each sample includes a raw video sequence, a natural language instruction, a multi-valued mask indicating the regions to be inpainted, and the corresponding inpainting result.

ROVI is built upon the YouTube-VOS Xu et al. (2018) and A2D-Sentences Xu et al. (2015) datasets, covering various scenarios such as dynamic object motion, varying backgrounds, and multiple scene types. The inclusion of both visual and textual information allows for comprehensive evaluation of language-guided inpainting models, particularly in assessing temporal consistency and mask-guided reconstruction accuracy.

Currently, ROVI is the only large-scale public benchmark available for this task; therefore, all our quantitative experiments are restricted to this dataset. While this setting provides a fair basis for comparison with prior work, further evaluations on additional benchmarks will be an important direction for future studies.

4.1.2 Training settings

All models are trained using the Adam optimizer with parameters $\beta_1 = 0$ and $\beta_2 = 0.99$, an initial learning rate of 1×10^{-4}. A MultiStepLR scheduler is employed with a milestone at 60,000 iterations and a decay factor of 0.1. Training is conducted for a total of 200,000 iterations. Video clips are loaded with 10 consecutive local frames and 6 reference frames per sample to provide sufficient temporal context. All frames are resized to a resolution of 432×240. Data loading is accelerated using 8 worker threads and 8 prefetch queues, with logging every 100 iterations and model checkpoints saved every 10,000 iterations.

Optical flow for temporal consistency is precomputed using the RAFT Teed & Deng (2020) model and stored in the specified flow directory. During training, these precomputed flows are loaded to reduce computational overhead. Mask post-processing is applied using a Gaussian blur with kernel size $(5, 5)$ and $\sigma = 0$, followed by binarization using a threshold of $\tau = 127$.

The overall training objective combines multiple loss components to balance reconstruction accuracy and temporal coherence. The masked region reconstruction loss λ_h is weighted by 6, the valid region reconstruction lossλ_v by 3, the feature warp lossλ_w by 0.5, and the Warp-Relation Consistency lossλ_r by 0.3. This configuration ensures that the model can effectively restore masked regions while preserving temporal coherence and minimizing flickering artifacts.

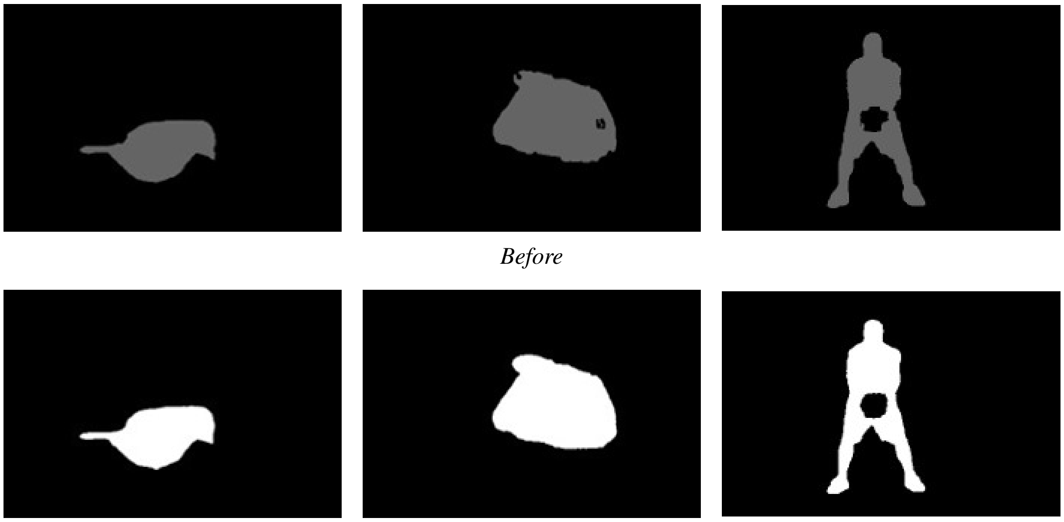


Figure 4: Comparison of masks before and after MSEM processing. The first row shows masks before processing, and the second row shows masks after processing.

4.1.3 Evaluation Metrics

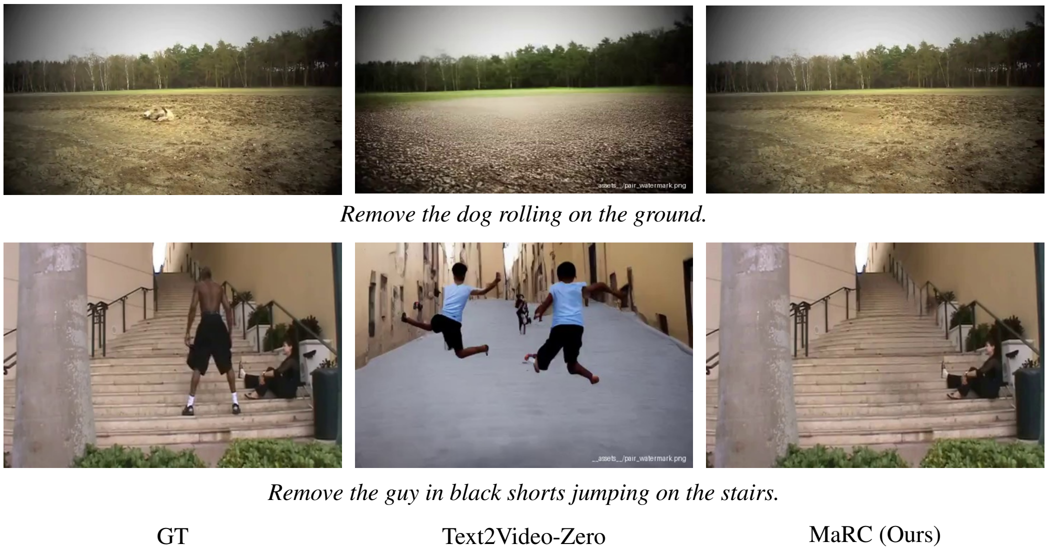
All experiments are conducted on a workstation equipped with two NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory, with the total training time approximately 3 days.

We evaluate our model from both image quality and temporal consistency perspectives.

For video quality, we use Peak Signal-to-Noise Ratio Hore & Ziou (2010) and Structural Similarity Index Measure Wang et al. (2004) . PSNR quantifies the pixel-wise reconstruction fidelity between the inpainted frames and the ground truth, with higher values indicating better reconstruction. SSIM measures perceptual similarity by considering luminance, contrast, and structural information, where higher scores correspond to more visually plausible results.

To assess temporal consistency, we employ the E_{warp} metric Lai et al. (2018), which measures the feature difference between adjacent frames after optical flow warping. Lower E_{warp} values indicate better temporal coherence, as the model is able to maintain consistent content across consecutive frames without introducing flickering artifacts.

By combining these metrics, we comprehensively evaluate the effectiveness of the proposed model in producing both high-quality frame-level reconstructions and temporally stable video sequences.


Remove the dog rolling on the ground.

Remove the guy in black shorts jumping on the stairs.

Figure 5: Comparison of video editing results. Each row corresponds to a different prompt, with columns showing GT, Text2Video-Zero, and MaRC (Ours).

4.2 Video inpainting results

4.2.1 Quantitative Comparison

We present qualitative comparisons to validate our approach. Fig. 2 shows inpainted results guided by prompts, while Fig. 3 provides side-by-side comparisons with baselines, where our method better preserves object structures and temporal consistency, whereas baselines often yield artifacts or incomplete restorations. Fig. 4 illustrates mask refinement in the MSEM module: the initial LISA masks are coarse with holes or jagged edges, and after Gaussian smoothing, thresholding, and morphological operations, the refined masks align more accurately with target regions and textual instructions. Fig. 5 compares our model with text-editing baselines, showing that while Text2Video-Zero removes occlusions, it often introduces unintended content. Overall, these results demonstrate that combining language-guided mask generation, MSEM refinement, and Warp-Relation Consistency yields visually plausible and temporally coherent inpainting.

4.2.2 Quantitative Comparison

We evaluate the performance of our model on the ROVI dataset using standard image quality and temporal consistency metrics, including PSNR, SSIM, and E_{warp}. As reported in Table 1, our approach consistently outperforms baseline methods across all metrics. Specifically, our model achieves higher PSNR and SSIM scores, indicating more accurate reconstruction of the missing

Table 1: Quantitative comparison on the ROVI dataset. Higher PSNR/SSIM indicate better spatial quality

Method	PSNR ↑	SSIM ↑
InstructPic2Pix	18.12	0.600
MagicBrush	20.39	0.725
Text2Video-Zero	19.58	0.671
LGVI	22.85	0.756
Ours	**28.71**	**0.915**

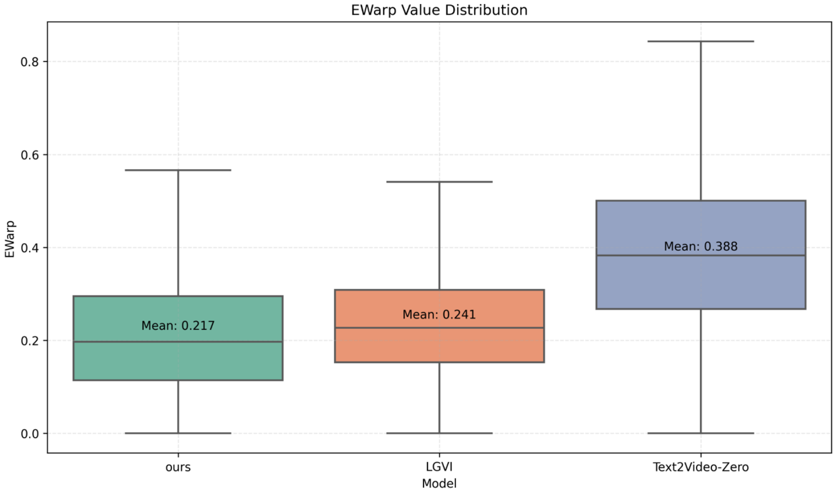


Figure 6: E_{warp} comparison with other models,Our model has the lowest mean and the best stability.

regions and better preservation of structural details. At the same time, the E_{warp} values are significantly lower, demonstrating improved temporal coherence and reduced flickering in consecutive frames.

The results also highlight the effectiveness of mask post-processing and the Warp-Relation Consistency loss. When these components are applied, both frame-level reconstruction and temporal consistency are further enhanced, suggesting that precise mask guidance and explicit temporal constraints are crucial for high-quality video inpainting. Overall, the quantitative comparisons confirm that our method achieves state-of-the-art performance in both spatial fidelity and temporal stability. we adopt a box plot visualization to provide a clearer view of the variance and robustness of each method across all test samples. As shown in Fig.6.

4.2.3 Ablating Design Components

To further verify the effectiveness of each proposed component, we conducted ablation experiments on the ROVI dataset. As shown in Table 2, starting from the baseline model, introducing the MSEM module can improve PSNR and SSIM. When Warp is introduced

When using the Warp-Relation Consistency, the E_{warp} score drops significantly. When we further merge the relationship-based modules, performance will be consistently enhanced and the best results will be achieved in all metrics. These results indicate that each component makes a positive contribution to the final performance, and their combination effectively enhances the reconstruction quality and temporal coherence.Numbers and Arrays

Table 2: Ablation experiment results. Comparison of each strategy

Warp-Relation	Mask	MSEM	PSNR ↑	SSIM ↑	E_{warp} ↓
×	✓	×	28.65	0.915	0.259
✓	✓	×	28.69	0.915	0.220
✓	×	×	12.35	0.436	0.349
✓	✓	✓	**28.71**	**0.915**	**0.218**

4.3 Conclusion

In this paper, we introduced a video inpainting framework driven by Multimodal Large Language Models. By combining language-guided mask generation with advanced restoration models, our method enables flexible text-driven editing. A morphology-based mask optimization reduces jagged edges and voids, while the proposed Warp-Relation Consistency enforces temporal alignment through optical flow and relation-aware constraints.

Experiments on public benchmarks show superior restoration fidelity and temporal coherence over state-of-the-art methods. Beyond performance gains, our work demonstrates the potential of MLLMs to bridge high-level semantic guidance with low-level restoration. Future directions include applying the framework to long-form videos and interactive editing for broader use in film, content creation, and immersive media.

A current limitation of our study is that experiments are conducted only on the ROVI dataset, due to the scarcity of large-scale public benchmarks for text-driven video inpainting. In future work, we plan to extend evaluation to additional datasets and more diverse scenarios, including higher-resolution and longer video sequences, to further validate the generalizability of our framework.+/-//

References

Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pp. I–I. IEEE, 2001.

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.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Alexandru Buburuzan, Anuj Sharma, John Redford, Puneet K Dokania, and Romain Mueller. Mobi: Multimodal object inpainting using diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1974–1984, 2025.

H. Chen, J. Ren, J. Gu, H. Wu, X. Lu, H. Cai, and L. Zhu. Snow removal in video: A new dataset and a novel method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13165–13176, 2023.

Suhwan Cho, Seoung Wug Oh, Sangyoun Lee, and Joon-Young Lee. Elevating flow-guided video inpainting with reference generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2527–2535, 2025.

Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17622–17631, 2022.

Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision*, pp. 713–729. Springer, 2020.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

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*, pp. 15954–15964, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 170–185, 2018.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Eunhye Lee, Jinsu Yoo, Yunjeong Yang, Sungyong Baik, and Tae Hyun Kim. Semantic-aware dynamic parameter for video inpainting transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12949–12958, 2023.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022.

Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, Shunli Zhang, and Robby T Tan. Nightrain: Nighttime video deraining via adaptive-rain-removal and adaptive-correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3378–3385, 2024.

Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14040–14049, 2021.

Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9117–9125, 2023.

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.

Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3511–3520, 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.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, et al. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2578–2588, 2025.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7623–7633, 2023.

Jianzong Wu, Xiangtai Li, Chenyang Si, Shangchen Zhou, Jingkang Yang, Jiangning Zhang, Yining Li, Kai Chen, Yunhai Tong, Ziwei Liu, et al. Towards language-driven video inpainting via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12501–12511, 2024a.

Zhiliang Wu, Changchang Sun, Hanyu Xuan, Gaowen Liu, and Yan Yan. Waveformer: wavelet transformer for noise-robust video inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6180–6188, 2024b.

C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 585–601, 2018.

Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3723–3732, 2019.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8703–8712, 2024.

Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European conference on computer vision*, pp. 528–543. Springer, 2020.

Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European conference on computer vision*, pp. 74–90. Springer, 2022.

Kaidong Zhang, Jialun Peng, Jingjing Fu, and Dong Liu. Exploiting optical flow guidance for transformer-based video inpainting. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):4977–4992, 2024.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Rong Xiao, Kam-Fai Wong, and Lei Zhang. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11067–11076, 2025.