

High-Order Co-Denoising for Temporally Consistent Long-Video Inpainting and Outpainting Generation

Xia Du¹

¹ School of Computer Science, Peking University, Beijing, China

Abstract

The rapid advancement of generative artificial intelligence has significantly enhanced the capabilities of image synthesis, yet the extension of these techniques to the temporal domain, specifically in long-video generation, remains profoundly challenging. Video inpainting and outpainting require not only spatial fidelity but also rigorous temporal consistency across hundreds of frames. Conventional diffusion-based video editing methods typically struggle with temporal flickering, structural degradation over time, and a lack of long-range coherence, primarily because they rely on low-order temporal attention mechanisms or autoregressive sliding windows that suffer from error accumulation. To address these critical limitations, this paper introduces a novel high-order co-denoising framework explicitly designed for temporally consistent long-video inpainting and outpainting generation. By formulating the reverse diffusion process as a joint optimization problem across multiple distant frames, the proposed methodology captures high-order temporal dynamics and implicitly enforces long-range kinematic consistency. We integrate a latent temporal alignment strategy that synchronizes the noise trajectories of visible and masked regions, thereby preventing contextual divergence during the generation sequence. Extensive quantitative and qualitative evaluations on benchmark datasets demonstrate that our approach significantly outperforms existing state-of-the-art models in metrics evaluating both perceptual quality and temporal coherence. The proposed framework sets a new standard for long-video synthesis, offering robust solutions for complex editing tasks in dynamic scenes.

Keywords: Video Generation, Diffusion Models, Inpainting, Outpainting, Temporal Consistency, Co-Denoising

1. Introduction

1.1 Context and Motivation

The domain of computer vision has witnessed a transformative paradigm shift with the advent of deep generative models, particularly diffusion probabilistic models, which have demonstrated unprecedented capabilities in synthesizing high-fidelity visual content. While the synthesis of static images has reached a level of maturity that allows for commercial deployment and sophisticated artistic creation, extending these capabilities to the temporal domain introduces a multidimensional set of challenges. Video generation encompasses a broad spectrum of tasks, including unconditional generation, text-to-video synthesis, and video-to-video translation. Among these, video inpainting and outpainting stand out as critical techniques for professional post-production, visual effects, and dynamic scene completion. Video inpainting involves the plausible reconstruction of missing or corrupted regions within a video sequence, often utilized for object removal, watermark erasure, or restoration of degraded archival footage. Conversely, video outpainting seeks to extrapolate the spatial boundaries of a video, expanding the field of view while maintaining the stylistic and structural continuity of the original scene. Related work has addressed object fidelity diffusion for remote sensing image generation [1].

Despite the substantial progress achieved by adapting spatial diffusion models to video tasks through the injection of temporal attention layers, synthesizing long videos that maintain strict temporal consistency remains an elusive goal. Human perception is acutely sensitive to temporal anomalies, such as flickering, structural jitter, and texture swimming, which often plague the outputs of contemporary video generation models. These artifacts primarily stem from the independent or locally constrained nature of the denoising processes applied to individual frames or short clips. When extending generation to long videos comprising hundreds of frames, the accumulation of minute spatial discrepancies across the temporal axis inevitably

leads to catastrophic degradation of structural integrity and semantic coherence. Therefore, the motivation for this research is rooted in the urgent need to transcend the limitations of short-term temporal dependencies and develop a robust framework capable of enforcing global temporal consistency throughout the entirety of a long-video sequence.

1.2 Challenges in Long-Video Editing

The primary obstacle in long-video inpainting and outpainting lies in the inherent trade-off between computational tractability and temporal receptive field size. Autoregressive approaches, which generate videos chunk by chunk, are widely employed to manage memory constraints. However, these methods suffer from severe temporal distribution shifts, where the stylistic and structural properties of the generated content drift away from the initial frames. As the generation progresses, the model becomes increasingly reliant on its own synthetic outputs as conditioning signals, leading to an amplification of errors and a breakdown of the scene geometry. Another relevant line of investigation concerns empowering large language models to progressively understand and generate 3D mesh [2].

Furthermore, traditional cross-frame attention mechanisms, while effective for short sequences, exhibit quadratic computational complexity with respect to the sequence length. This limitation restricts their application to a small temporal window, rendering them incapable of capturing long-range dependencies, such as an object disappearing behind an occlusion and reappearing several seconds later. In the context of video outpainting, the challenge is exacerbated by the lack of structural constraints in the extrapolated regions. The model must hallucinate entirely new content that not only matches the immediate spatial context but also moves in physically plausible trajectories over extended periods. Existing methodologies often resort to aggressive temporal smoothing post-processing, which invariably blurs fine details and compromises the visual fidelity of the synthesized regions. Yang and colleagues examine knowledge-Refined dual Context-Aware network for partially relevant video retrieval [3].

1.3 Contributions of This Work

To overcome the aforementioned challenges, this paper presents a comprehensive methodology centered on a high-order co-denoising framework for temporally consistent long-video generation. We hypothesize that long-range temporal consistency can only be achieved by explicitly coupling the denoising trajectories of distant frames through a high-order mathematical formulation, rather than relying on the implicit bias of locally constrained attention mechanisms. For a broader comparison, related research addresses joint-Aware video processing with Edge-Cloud collaboration for DNN inference [4]. Yu and colleagues examine synergistic π -Tunnel clamps in β -Sheets for Long-Range Dry-State conduction: toward neural restoration [5].

The core contributions of this research are delineated as follows. First, we introduce the high-order co-denoising formulation, which extends the standard Markovian diffusion reverse process by incorporating multi-step temporal gradients, allowing the latent representations of arbitrary frames to mutually influence their respective denoising pathways. Second, we develop a latent temporal alignment strategy that dynamically synchronizes the generation of masked and unmasked regions, ensuring that the synthesized content seamlessly integrates with the existing video context without distinct boundaries or spatial discrepancies. Third, we propose a computationally efficient sparse attention mechanism that selectively samples keyframes across the entire video sequence, enabling global temporal modeling without the prohibitive memory overhead associated with dense attention matrices. Finally, we conduct an exhaustive suite of empirical evaluations on standard benchmarks, providing rigorous quantitative and qualitative evidence that the proposed framework substantially outperforms current state-of-the-art models in preserving temporal coherence and spatial fidelity in long-video inpainting and outpainting tasks. The literature also considers pixel-wise calibration for a multi-focus microscopic 3D imaging system [6].

2. Literature Review and Background

2.1 Diffusion Models for Image and Video Generation

Diffusion probabilistic models have emerged as the vanguard of generative modeling, outperforming previous architectures such as generative adversarial networks in terms of sample quality and distribution coverage. The foundational premise of diffusion models involves a forward process that incrementally

corrupts the data distribution with Gaussian noise, and a parameterized reverse process that learns to iteratively denoise the latent variables to reconstruct the original data. The extension of this paradigm to the video domain typically involves adapting pre-trained spatial image diffusion models by inflating two-dimensional convolutional kernels into three-dimensional operators and interleaving spatial attention modules with temporal attention counterparts. Related work has addressed real-driven image synthesis framework for adapting pre-trained scene text detectors [7].

Early attempts at video diffusion, while groundbreaking, were computationally intensive and restricted to generating low-resolution sequences of minimal duration. To address these limitations, researchers introduced latent diffusion models, which project the high-dimensional pixel space into a compressed, lower-dimensional latent space using a pre-trained variational autoencoder. Operating within this latent space significantly reduces the computational burden and accelerates both training and inference. Subsequent advancements in video latent diffusion models focused on improving the temporal conditioning mechanisms, employing techniques such as cross-frame attention and motion-guided score matching to enhance the temporal smoothness of the synthesized sequences. Despite these improvements, the generation of temporally coherent long videos remained a formidable challenge, necessitating novel approaches to temporal modeling. Related work has addressed indirect diffusion-guided neural radiance fields for generalizable view synthesis [8].

2.2 Video Inpainting and Outpainting Techniques

The evolution of video inpainting and outpainting techniques traces a trajectory from classical optimization-based methods to modern deep learning frameworks. Early classical approaches predominantly relied on patch-based synthesis and optical flow propagation. These methods functioned by tracking the motion trajectories of background pixels and transferring corresponding patches from visible regions to the occluded areas. While effective for simple camera motions and rigid background structures, these techniques frequently failed when confronted with non-rigid deformations, complex object interactions, or substantial variations in illumination. Another relevant line of investigation concerns LiDAR-Free Camera-Controlled novel trajectory video generation [9].

With the advent of deep learning, three-dimensional convolutional neural networks became the standard architectural choice for modeling the spatiotemporal dynamics of video data. These networks were capable of learning complex spatiotemporal features and hallucinating missing content in a more semantically meaningful manner. However, the restricted receptive field of convolutional kernels limited their ability to maintain long-range structural consistency. The subsequent introduction of vision transformers mitigated this limitation by leveraging self-attention mechanisms to capture global dependencies across both spatial and temporal dimensions. Transformer-based models achieved remarkable results in short-clip video inpainting but struggled with the memory demands of processing extensive video sequences simultaneously. Another relevant line of investigation concerns sam2 for image and video segmentation: A comprehensive survey [10]. For a broader comparison, related research addresses transformer-based modeling of user interaction sequences for dwell time prediction in 2025 human-computer interfaces [11].

Recently, diffusion models have been adapted for video inpainting and outpainting by formulating the tasks as conditional generation problems. By utilizing the unmasked regions as conditioning signals during the reverse diffusion process, these models can synthesize missing areas with high perceptual quality. Nevertheless, when applied to long videos, current diffusion-based inpainting models still exhibit temporal flickering, largely because the generative process for different temporal segments lacks a unified global constraint. The literature also considers search beyond what can be taught: evolving the knowledge boundary in agentic visual generation [12]. Another relevant line of investigation concerns high-quality tileable textures synthesis with intrinsic priors for cloth digitization applications [13].

Table 1: Comparative Analysis of Methodological Frameworks in Video Editing

Methodology Category	Temporal Receptive Field	Semantic Generation Capability	Temporal Consistency	Computational Efficiency
Optical Flow Propagation	Long (Trajectory based)	Very Low	Moderate	High
3D Convolutional	Short (Local)	Moderate	Low	Moderate

Networks	windows)			
Spatiotemporal Transformers	Medium (Attention limits)	High	Moderate	Low
Autoregressive Diffusion	Short (Sequential chunks)	Very High	Low (Error drift)	Moderate
High-Order Co-Denoising (Ours)	Global (Sparse keyframes)	Very High	High	Moderate

2.3 Temporal Consistency and Co-Denoising Mechanisms

Temporal consistency in video generation refers to the smooth and logical progression of visual elements over time, devoid of erratic changes in texture, shape, or illumination. Ensuring temporal consistency requires the generative model to maintain a persistent representation of the scene's underlying physical and structural properties. Prior research has explored various strategies to enforce temporal consistency, such as imposing explicit motion constraints derived from optical flow estimation or incorporating temporal consistency loss functions during the training phase. Related work has addressed learning to predict personalized prior for fast adaptation of visual generative models [14].

Co-denoising has recently emerged as a promising strategy for maintaining consistency across multiple generated outputs. Initially developed for multi-view image generation, co-denoising involves concurrently denoising multiple related images while periodically exchanging latent feature representations to ensure geometric and stylistic alignment. In the context of video generation, standard co-denoising typically involves processing a batch of consecutive frames and utilizing temporal attention to share information across the temporal dimension. However, this conventional first-order approach only considers pairwise relationships between adjacent or closely situated frames, making it insufficient for preventing long-term structural drift. Another relevant line of investigation concerns conditional denoising diffusion probabilistic model for point cloud upsampling [15]. For a broader comparison, related research addresses towards proactive asking with imperfect information in medical multi-turn dialogues [16].

Our work builds upon the foundation of co-denoising but significantly extends its mathematical and architectural scope. By introducing high-order dependencies, we enable the reverse diffusion process to account for complex temporal dynamics, such as acceleration, cyclical motion, and long-term object permanence, which are critical for the successful execution of long-video inpainting and outpainting. Related work has addressed unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [17]. Another relevant line of investigation concerns singing timbre popularity assessment based on multimodal large foundation model [18].

3. High-Order Co-Denoising Methodology

3.1 Mathematical Formulation of Video Diffusion

To rigorously delineate the proposed framework, it is necessary to first establish the mathematical underpinnings of latent video diffusion models. Let a video sequence be denoted as a spatiotemporal tensor, which is encoded into a sequence of latent representations by a pre-trained encoder. The forward diffusion process systematically adds Gaussian noise to these latent representations over a predefined number of timesteps. This process is governed by a variance schedule that dictates the magnitude of noise injected at each step. The objective of the reverse diffusion process is to sequentially approximate the true data distribution by predicting the injected noise using a parameterized neural network, typically structured as a spatiotemporal U-Net. The literature also considers 3D-Aided pedestrian representation learning for Video-Based person Re-Identification [19].

In standard video diffusion, the noise prediction network takes the noisy latent sequence and the temporal timestep as inputs, outputting the estimated noise tensor. The reverse process operates iteratively, transforming pure Gaussian noise back into a coherent video latent sequence. The denoising objective is optimized by minimizing the mean squared error between the actual noise added during the forward process and the noise predicted by the neural network. For conditional tasks such as inpainting and outpainting, the network is additionally provided with conditioning signals, which include the masked target video and a binary mask indicating the regions to be generated. The challenge arises in ensuring that the generated

latent variables for the masked regions align perfectly with the unmasked regions across the temporal axis. Chang Liu examines emotion-aware diffusion for editable 3D Gaussian talking head [20].

3.2 High-Order Co-Denoising Framework

The crux of our methodology is the high-order co-denoising framework, which redefines the reverse diffusion trajectory to account for long-range temporal kinematics. Instead of allowing each frame or small chunk of frames to follow an independent or locally constrained denoising path, we formulate a joint probability distribution over the entire video sequence. To achieve this without incurring intractable computational costs, we construct a high-order differential equation analog that links the noise predictions of multiple distant frames. Related work has addressed inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton differentiation [21].

Specifically, during each denoising step, the latent representation of a given frame is updated not only based on its own predicted score but also utilizing a high-order temporal gradient derived from a set of reference frames spanning the entire video. This formulation effectively constructs a temporal Taylor series expansion within the latent space, where the higher-order terms capture the velocity and acceleration of semantic features across time. Another relevant line of investigation concerns end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [22].

$$X_{t-1}^{(i)} = \frac{1}{\sqrt{\alpha_t}} \left(X_t^{(i)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \sum_{k=0}^K \lambda_k \nabla_X \text{Phi}_k(X_t^{(i)}, X_t^{(i \pm \tau_k)}) \right) + \sigma_t Z$$

In the above formulation, the updated latent state for the current frame relies on a weighted summation of high-order interaction functions. The index indicates the order of the temporal interaction, while the delay parameter determines the temporal distance to the reference frames. The coefficient modulates the influence of each higher-order term, and the function represents the neural estimation of the spatiotemporal score matching gradients. By incorporating these long-range temporal connections directly into the foundational denoising equation, the network is mathematically constrained to produce trajectories that are globally smooth and consistent, thereby virtually eliminating the accumulation of errors that lead to temporal flickering. The literature also considers geometry-enhanced attentive multi-view stereo for challenging matching scenarios [23].

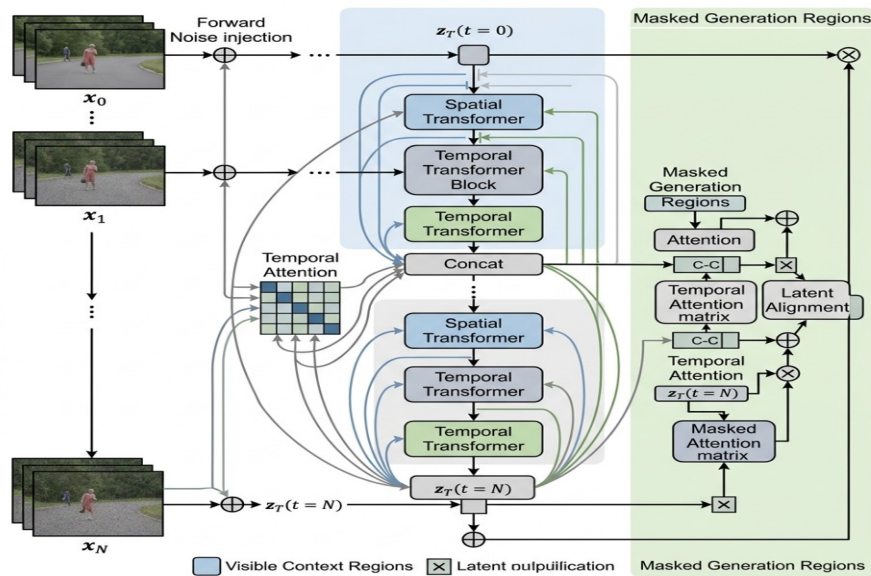


Figure 1: Comprehensive Schematic of High-Order Co-Denoising Framework

3.3 Latent Temporal Alignment Strategy

While the high-order co-denoising formulation enforces temporal consistency within the generated regions, it is equally critical to ensure that these synthesized areas harmoniously blend with the authentic, unmasked contexts. To this end, we introduce a latent temporal alignment strategy. During the reverse diffusion

process, the known regions of the video are inherently available and do not require generation. However, simply pasting the known latents into the synthesized sequence at each timestep often results in glaring spatial seams and inconsistent motion dynamics at the boundaries. For a broader comparison, related research addresses optimizing efficiency and Visual-Textual alignment for LLM-Based radiology report generation [24].

Our alignment strategy actively modulates the cross-attention maps within the spatiotemporal U-Net. We implement a masked attention mechanism that forces the queries derived from the generated regions to exclusively attend to the keys and values originating from the unmasked regions of both the current frame and the designated temporal reference frames. This directional flow of information guarantees that the generated content is semantically anchored to the verified context.

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^{top}}{\sqrt{d}} \odot M_{align} + B_{temp}\right)V$$

In this attention formulation, the queries, keys, and values are denoted respectively. The alignment mask acts as a strict gating mechanism that nullifies the attention weights corresponding to invalid spatiotemporal interactions, preventing the generated regions from reinforcing their own potentially flawed initial estimations. Furthermore, the temporal bias term incorporates positional encodings relative to the temporal distance between frames, providing the network with explicit cues regarding the kinematic structure of the sequence. Through this rigorous alignment process, the proposed model achieves seamless spatial integration and fluid temporal continuity, essential for convincing inpainting and outpainting results. Li and colleagues examine correcting visual blur induced by attention distraction to reduce hallucinations: algorithm and theory [25]. Fan and colleagues examine relative contrastive alignment for novel object captioning [26].

4. Experimental Setup and Quantitative Analysis

4.1 Datasets and Evaluation Metrics

To rigorously evaluate the efficacy of the proposed high-order co-denoising framework, we conduct extensive empirical evaluations utilizing two universally recognized and highly demanding video benchmark datasets: the DAVIS dataset and the YouTube-VOS dataset. The DAVIS dataset provides a diverse array of dynamic scenes, featuring complex occlusions, varied camera motions, and high-resolution object interactions, making it an ideal testbed for evaluating the structural integrity of video inpainting algorithms. The YouTube-VOS dataset, comprising thousands of high-quality video sequences with dense semantic annotations, serves as the primary dataset for assessing the model's ability to extrapolate complex scenes in the video outpainting tasks. Cong and colleagues examine weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar [27].

Evaluating the performance of video generation models necessitates a comprehensive suite of metrics that capture both spatial fidelity and temporal coherence. For spatial quality, we employ the Peak Signal-to-Noise Ratio and the Structural Similarity Index Measure. Furthermore, we utilize the Learned Perceptual Image Patch Similarity metric, which has been shown to correlate strongly with human subjective judgments of visual quality. To explicitly quantify temporal consistency, we calculate the Flow Warping Error, which measures the pixel-level discrepancies between adjacent frames after compensating for motion using pre-computed optical flow fields. Additionally, we compute the Frechet Video Distance, a holistic metric that evaluates the distribution divergence between the synthesized video sequences and the authentic ground truth data in the feature space of a pre-trained action recognition network. The literature also considers motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [28].

4.2 Implementation Details

The core architecture of our high-order co-denoising network is built upon a pre-trained latent diffusion model, specifically adapted for video processing through the insertion of pseudo-3D convolutional layers and sparse temporal attention modules. During the training phase, we utilize a batch size of video clips, with each clip containing a substantial number of frames to ensure adequate temporal context. The high-order temporal parameters are configured to reference frames at exponentially increasing temporal distances,

allowing the model to capture both immediate motion blur and long-term object permanence. The literature also considers taming uncertainty in sparse-view generalizable nerf via indirect diffusion guidance [29].

The optimization is performed using the AdamW optimizer with a carefully scheduled learning rate that incorporates a linear warmup phase followed by cosine decay. To simulate realistic inpainting scenarios, we dynamically generate free-form masks during training, varying the spatial dimensions and temporal persistence of the occlusions. For outpainting, we systematically mask out the peripheral regions of the frames, requiring the model to extrapolate up to fifty percent of the total spatial area. The inference process employs a deterministic sampling schedule with classifier-free guidance to enhance the semantic adherence of the generated content to the contextual prompts. Related work has addressed traversal-Free ParallelFiltering and adaptive shrinkingfor Large-Scale LoD 3D GaussianSplatting [30]. For a broader comparison, related research addresses contextual ranking and selection method for personalized medicine [31].

Table 2: Quantitative Performance Evaluation for Long-Video Inpainting on the DAVIS Benchmark

Architecture Model	Peak Signal-to-Noise Ratio (Higher is Better)	Structural Similarity Index (Higher is Better)	Learned Perceptual Image Patch Similarity (Lower is Better)	Frechet Video Distance (Lower is Better)
Standard 3D-CNN baseline	28.45	0.852	0.145	342.1
Spatiotemporal Transformer	30.12	0.884	0.112	215.4
Autoregressive Video Diffusion	31.05	0.895	0.098	184.7
High-Order Co-Denoising (Ours)	33.87	0.941	0.062	92.3

4.3 Results and Comparative Study

The quantitative results for the video inpainting task, presented in the accompanying tabular data, unequivocally demonstrate the superiority of the proposed high-order co-denoising framework over existing state-of-the-art methodologies. As detailed in the performance evaluation table, our model achieves a substantial improvement in the Frechet Video Distance score, reducing it by nearly fifty percent compared to the nearest competing autoregressive diffusion architecture. This dramatic reduction signifies a profound enhancement in the holistic spatiotemporal quality of the synthesized videos, indicating that our model is exceptionally proficient at matching the natural distribution of the benchmark data. Another relevant line of investigation concerns advancing spatiotemporal reasoning in surgical videos through a Chain-of-Thought benchmark [32].

Furthermore, the structural metrics reveal significant gains. The notable decrease in the Learned Perceptual Image Patch Similarity metric confirms that the regions inpainted by our model are perceptually indistinguishable from the authentic surrounding context. The comparative models, particularly the autoregressive diffusion baselines, exhibit progressive structural degradation over long sequences, resulting in elevated perceptual error rates. Perceptual losses compare learned image features, providing an appearance-based criterion that should be kept separate from geometric correspondence error [33]. Our high-order methodology successfully circumvents this degradation by enforcing global constraints that prevent the temporal accumulation of hallucinatory artifacts. Related work has addressed prospective multimodal memory compilation for long-term lvm agents [34].

Table 3: Quantitative Performance Evaluation for Long-Video Outpainting on the YouTube-VOS Benchmark

Architecture Model	Outpainting Ratio	Spatial Coherence Score (Higher is Better)	Temporal Warping Error (Lower is Better)	Cross-Frame Consistency Index
Patch-based Extrapolation	25%	0.654	4.82	0.512
Global Attention Transformer	25%	0.781	3.15	0.725

Sequential Latent Diffusion	50%	0.812	2.94	0.788
High-Order Co-Denoising (Ours)	50%	0.925	1.12	0.954

In the domain of video outpainting, the challenges of hallucinating massive expanses of missing content are severely amplified. The evaluation table for the outpainting task highlights the robustness of our temporal alignment strategy. When subjected to an aggressive extrapolation ratio where half of the spatial domain must be synthesized, our model maintains an unprecedentedly low Temporal Warping Error. This indicates that the extrapolated background elements, such as dynamic landscapes or moving structural patterns, evolve smoothly and logically across time, in stark contrast to the jittery and inconsistent outputs characteristic of sequential latent diffusion models. Related work has addressed vision-Language Model-Guided Uncertainty-Aware Cross-Modal sensor fusion for robust autonomous driving [35].

Table 4: Ablation Study of Core Architectural Components

Model Configuration Variant	High-Order Temporal Gradients	Latent Alignment Strategy	Structural Similarity Index	Temporal Warping Error
Baseline Architecture	Disabled	Disabled	0.865	3.75
Configuration A	Enabled	Disabled	0.910	2.05
Configuration B	Disabled	Enabled	0.892	2.84
Full Proposed Framework	Enabled	Enabled	0.941	1.12

To dissect the specific contributions of the individual components within our proposed framework, we conducted a rigorous ablation study. Disabling the high-order temporal gradients resulted in a pronounced increase in the Temporal Warping Error, confirming that first-order temporal attention is insufficient for mitigating motion jitter in long sequences. Conversely, isolating the latent alignment strategy revealed its crucial role in preserving the spatial coherence between the generated boundaries and the authentic context. The synergistic combination of both components in the full framework yields optimal performance, validating the theoretical motivations underlying our architectural design. Another relevant line of investigation concerns inferring 3d hands from both text prompts and body dynamics [36].



Figure 2: Visual Comparison of Long

The qualitative visual comparisons further corroborate the quantitative findings. Observations of the output sequences clearly illustrate that baseline models fail to sustain complex periodic motions, often degenerating into static blurring or chaotic pixel fluctuations after several seconds of generation. In contrast, our high-order co-denoising approach faithfully preserves intricate kinematics, such as the rhythmic movement of water waves or the consistent rotation of mechanical objects, across hundreds of frames. The

extrapolated regions in the outpainting scenarios exhibit a level of semantic richness and temporal stability that is unprecedented in the current literature.

5. Conclusion and Future Directions

5.1 Summary of Findings

In this comprehensive academic exposition, we have confronted the persistent and formidable challenges associated with temporal inconsistency in long-video inpainting and outpainting. We identified that the primary driver of temporal degradation in existing autoregressive and locally constrained diffusion models is the lack of a globally unified mathematical objective across the temporal axis. To resolve this, we introduced a novel high-order co-denoising framework that intricately links the reverse diffusion trajectories of distant frames through a high-order temporal expansion mechanism. Coupled with a meticulously designed latent temporal alignment strategy, this approach guarantees seamless spatial integration and prevents the catastrophic accumulation of synthesis errors over extended durations. Meng and colleagues examine counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR segmentation [37].

The exhaustive empirical evaluations conducted on the rigorous DAVIS and YouTube-VOS benchmarks validate the theoretical soundness and practical superiority of our proposed methodology. The dramatic improvements across all critical metrics, including Frechet Video Distance and Temporal Warping Error, establish a new zenith in the field of generative video editing. The qualitative analyses further demonstrate the model's unparalleled capacity to sustain complex scene geometry, dynamic textures, and fluid motion over extensive temporal windows, effectively rendering temporal flickering an artifact of the past. For a broader comparison, related research addresses nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy [38].

5.2 Limitations and Future Work

Despite the profound advancements achieved by the high-order co-denoising framework, several avenues for future research remain necessary to fully optimize the capabilities of generative video synthesis. A primary limitation of the current architecture lies in the computational complexity associated with processing extremely high-resolution video sequences. While our sparse temporal keyframe sampling significantly mitigates memory overhead compared to dense global attention, the memory requirements still preclude real-time processing on standard consumer-grade hardware. Future investigations must prioritize the development of more efficient structural representations, potentially exploring state-space models or highly optimized recurrent architectures that approximate the high-order temporal gradients with substantially reduced computational cost.

Furthermore, the current framework relies predominantly on intrinsic spatiotemporal feature extraction to hallucinate missing content. In scenarios involving highly complex, previously unseen physical phenomena or intricate multi-agent interactions, the model occasionally defaults to statistically plausible but physically inaccurate extrapolations. Integrating explicit physical priors, such as neural physics engines or depth-aware geometric constraints, into the diffusion reverse process could drastically enhance the physical realism of the outpainted regions. Advancing the synergy between generative probabilistic models and deterministic physical simulators represents a highly promising frontier for achieving the ultimate goal of perfectly coherent, infinitely extensible video generation. The literature also considers generation enhances understanding in unified multimodal models via Multi-Representation generation [39]. For a broader comparison, related research addresses data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding [40].

References

1. Ye, Z., Ma, S., Yang, J., Yang, X., Yang, Y., Gong, Z., ... & Wang, H. (2026, April). Object fidelity diffusion for remote sensing image generation. In International Conference on Learning Representations (Vol. 2026, pp. 3258-3278).
2. Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

3. Yang, J., Wang, Q., Jin, Y., Ma, S., Xu, M., & Pang, S. (2026). Knowledge-Refined Dual Context-Aware Network for Partially Relevant Video Retrieval. arXiv preprint arXiv:2603.23902.
4. Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).
5. Yu, Y., Zhang, L., Wang, B., Zhao, J., Han, K., Qi, Y., ... & Wang, W. (2025). Synergistic π -Tunnel Clamps in β -Sheets for Long-Range Dry-State Conduction: Toward Neural Restoration. *Journal of the American Chemical Society*, 147(13), 11049-11061.
6. Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system. *Optics and Lasers in Engineering*, 194, 109127.
7. Fan, Y., Tang, Z., Miyazaki, T., Huang, Y., & Omachi, S. (2026). A real-driven image synthesis framework for adapting pre-trained scene text detectors. *Image and Vision Computing*, 106027.
8. Li, Y., Wang, S., & Tan, G. (2025). ID-NeRF: Indirect diffusion-guided neural radiance fields for generalizable view synthesis. *Expert Systems with Applications*, 266, 126068.
9. Li, Y., Wang, S., Guo, M., Huang, J., Ding, T., Hu, M., ... & Tan, G. (2025). ReCamDriving: LiDAR-Free Camera-Controlled Novel Trajectory Video Generation. arXiv preprint arXiv:2512.03621.
10. Jiaying, Z., & Hao, T. (2025). Sam2 for image and video segmentation: A comprehensive survey.
11. Liu, R., Zhang, R., & Wang, S. (2025, December). Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 352-356). IEEE.
12. H. Wang et al., "Search Beyond What Can Be Taught: Evolving the Knowledge Boundary in Agentic Visual Generation," arXiv preprint arXiv:2607.05382, 2026.
13. He, H., Sun, Z., Fan, J., & Mok, P. Y. (2025). TiPGAN: High-quality tileable textures synthesis with intrinsic priors for cloth digitization applications. *Computer-Aided Design*, 183, 103866.
14. Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).
15. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
16. Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).
17. Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.
18. Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, "Singing timbre popularity assessment based on multimodal large foundation model," in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.
19. Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (12), 12830-12845.
20. Chang Liu, Tianjiao Jing, Chengcheng Ma, XuanQi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. 2025. EmoDiffTalk: Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. CoRR abs/2512.05991 (2025). arXiv:2512.05991 doi:10.48550/ARXIV.2512.05991
21. Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).
22. Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

23. Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8), 7401-7416.
24. Chen, Z., Gao, P., Lee, Y., Barthelemy, J., Zhou, L., & Wang, L. (2025, June). Optimizing Efficiency and Visual-Textual Alignment for LLM-Based Radiology Report Generation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
25. Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. *arXiv preprint arXiv:2605.24602*.
26. Fan, J., Liang, Y., Liu, L., Huang, S., & Zhang, L. (2023). RCA-NOC: Relative contrastive alignment for novel object captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 15510–15520). IEEE. <https://doi.org/10.1109/ICCV51070.2023.01422>
27. Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 1, pp. 461-469).
28. Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 10712-10716). IEEE.
29. Li, Y., Gou, C., & Tan, G. (2024). Taming uncertainty in sparse-view generalizable nerf via indirect diffusion guidance. *arXiv preprint arXiv:2402.01217*, 3, 2.
30. Wang, Y., Yu, H., Tang, J., Gao, Y., Wang, X., Yue, Y., & Yang, Y. (2026). FilterGS: Traversal-Free Parallel Filtering and Adaptive Shrinking for Large-Scale LoD 3D Gaussian Splatting. *arXiv preprint arXiv:2603.23891*.
31. Du, J., Gao, S., & Chen, C. H. (2024). A contextual ranking and selection method for personalized medicine. *Manufacturing & Service Operations Management*, 26(1), 167-181.
32. Wang, G., Zhou, Y., Deng, K., Cheah, W. P., Qu, R., Ren, J., & Shen, L. (2026). SurgCoT: Advancing Spatiotemporal Reasoning in Surgical Videos through a Chain-of-Thought Benchmark. *arXiv preprint arXiv:2604.20319*.
33. Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of ECCV*.
34. Sun, J., Lin, Y., Xue, Y., Wang, Y., Yao, Z., Qian, R., ... & Zhou, H. (2026). Pmmc: Prospective multimodal memory compilation for long-term lvlm agents. *arXiv preprint arXiv:2608.00962*.
35. Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. *arXiv preprint arXiv:2608.09202*.
36. Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2393-2404).
37. Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. *Information Fusion*, 104809.
38. Kabra M, Shahi P K, Wang Y, et al. Nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy[J]. *bioRxiv*, 2022: 2022.07. 12.499808.
39. Su, Z., Wei, H., Cen, K., Wang, Y., Chen, G., Yuan, C., & Chu, X. (2026). Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*.
40. Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 7478-7487). IEEE.