DiffMVR: An Intelligent Diffusion-based Multi-Guidance System for Real-time Video Restoration

Zheyan Zhang[a], Diego Klabjan[a], Renee CB Manworren[b]

[a]Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Rd, Evanston, IL 60208, United States, [b]College of Nursing and Health Innovation, University of Texas at Arlington, 411 S Nedderman Dr, Arlington, TX 76019, United States.

ABSTRACT

This paper introduces DiffMVR, a novel automated diffusion-based video restoration system designed to reconstruct occluded regions within dynamic, real-world settings, such as facial restoration in healthcare environments. In these environments, continuous, precise visual observation is critical yet often compromised. To address this, DiffMVR leverages dual adaptive reference frames as dynamic guidance, enabling real-time adaptability to complex visual contexts. Our framework employs parallel attention pipelines to intelligently integrate spatial and temporal features, guided by a hybrid loss function that simultaneously optimizes denoising accuracy and motion coherence, enabling precise control over the inpainting direction while preserving fine-grained details and ensuring smooth temporal transitions. Extensive experiments demonstrate that DiffMVR significantly outperforms current state-of-the-art methods across key metrics, including SSIM, FID, FVD, and Temporal Consistency, especially excelling in complex, rapidly changing visual contexts. Beyond quantitative gains, DiffMVR substantially advances downstream restoration tasks, particularly in reconstructing realistic and accurate facial features vital for neonatal care. Moreover, DiffMVR offers practical guidelines for integration within existing clinical monitoring platforms, thereby boosting both operational efficiency and accuracy. These advantages underscore the model's potential as an intelligent system for robust, real-time, and context-aware video restoration in dynamic and challenging scenarios.

Key words: Facial restoration, Clinical monitoring, Context-aware algorithms, Video inpainting, Diffusion models

1 BACKGROUND AND SIGNIFICANCE

The rise of diffusion models has revolutionized computer vision, driving advances in image editing (Kim et al., 2022), super-resolution (Hsu et al., 2024), object removal (Xu et al., 2023), colorization (Zhang et al., 2016), and restoration (Liang et al., 2021). While image-based inpainting methods excel at restoring missing regions in static frames, they lack the temporal coherence required for video-level tasks.

Building on image-based models, recent advancements in video-level inpainting have targeted the reconstruction of missing or occluded regions in sequence-level challenges that 2D image inpainting alone cannot fully overcome. Advances in deep learning have driven substantial progress in video inpainting. For example, Ouyang et al. (2021) utilize the convolutional neural networks for video inpainting, preserving high-frequency details. Further advancements, such as the First Frame Filling video inpainting model (Lee et al., 2025), leverage diffusion models to achieve accurate object removal, even with large masks. More recent models, including the Any-length video inpainting model (Zhang et al., 2024) and MotionAura (Susladkar et al., 2025), introduce diffusion-based video inpainting frameworks that support various video lengths. However, existing models often prioritize temporal consistency or flexibility over photorealistic quality, or concentrate on removing objects from videos rather than replacing them with precise, detailed alternatives. This leaves a significant gap in applications requiring high-fidelity content reconstruction.

To bridge this divide between temporal coherence and naturalistic fidelity, we introduce DiffMVR, a novel diffusion-based framework for dynamic, pairwise image-guided video inpainting. Our approach introduces two adaptive guiding images to steer inpainting precisely across detailed and complex video sequences. To the best of our knowledge, this is the first work to explore dual-image-guided video inpainting, a crucial advancement for tasks such as restoring facial movements where both faithful reconstruction and global continuity are essential. Our method fills the gap identified in current video inpainting approaches by combining a dual-image guidance mechanism with a novel motion loss term. A further discussion of our method and motivation is presented in Appendix *Motivation*.

The contributions of our approach are fourfold. (1) We tackle the often-overlooked core challenge of truly high-quality inpainting. Our systematic approach excels at accurately reconstructing subtle motions and fine facial details, even under challenging conditions. By guiding the diffusion process with dual images instead of relying solely on text prompts, our method captures detailed visual cues and provides

enhanced control over the inpainting process compared to text-based approaches. Moreover, our fully automated and user-friendly pipeline facilitates seamless integration into a variety of downstream tasks. (2) We propose a novel architecture that processes the symmetric and past unobstructed frames in parallel attention pipelines, intelligently fusing their outputs within the U-Net to provide comprehensive spatio-temporal guidance. (3) We introduce a hybrid loss function, nested within the diffusion process, that merges denoising and motion-consistency terms, enabling effective feature extraction from both the current and neighboring frames. Our method harnesses the power of diffusion for progressive frame restoration and optimizes the interaction between structural and temporal data, setting a new standard for precision in video inpainting. (4) Through quantitative and qualitative comparisons, we demonstrate that DiffMVR consistently outperforms state-of-the-art inpainting models. This work paves the way for more robust and reliable AI-driven video inpainting, improving decision-making in real-world scenarios.

2 RELATED WORK

Images are a crucial medium for information dissemination, but they are often susceptible to noise, damage, and interference, which can impede data analysis and knowledge extraction. To restore damaged images and design images according to human intent, various image inpainting approaches have emerged in recent years.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) initiated the development of AI-based image editing (a relevant review of GANs is given in Appendix *Extended Literature Review*), but their training instability and reliance on large datasets limit their effectiveness.

To mitigate these limitations, Pathak et al. (2016) introduced the encoder-decoder structure, providing a more stable approach for filling missing regions. More recently, diffusion-based models such as DDPM (Ho et al., 2020) and DDIM (Song et al., 2021) have been proposed to resolve issues like mode collapse and handle complex distributions, resulting in high-quality inpainted images with enhanced stability.

Diffusion models initially struggle to learn effectively from unmasked surrounding pixels (Saharia et al., 2022). To overcome this, text-guided models have been introduced to provide finer control over the inpainting process (Rombacj et al., 2022). Blended Diffusion (Avrahami et al., 2022), for example, integrates a pretrained CLIP model with DDPM to perform localized edits by encouraging the output to align with the text prompt using the CLIP score. Similarly, GLIDE (Nichol et al., 2022) and ContextDiff (Yang et al., 2024)

enhance semantic consistency, the former using CLIP with classifier-free guidance, while the latter incorporates cross-modal context into both forward and reverse diffusion processes.

As image inpainting techniques have matured, their extension to videos has followed naturally, motivated by applications in remote sensing, medical imaging, and traffic analysis (see Appendix *Extended Literature Review* for relevant works). However, video inpainting faces additional challenges, including high computational costs and difficulties in maintaining temporal consistency, often leading to motion artifacts.

Recently, researchers have been actively investigating a range of approaches to minimize the limitations imposed by these challenges, and diffusion-based methods represent one promising direction. General methods like DiffuEraser (Wu et al., 2023) leverage stable diffusion by incorporating prior information and expanding temporal receptive fields, effectively mitigating noisy artifacts. Guided techniques, exemplified by AVID (Zhang et al., 2024), employ consistent text prompts to steer object removal and ensure semantic alignment across frames. Complementing these approaches, transformer and propagation-based models such as Propainter (Zhou et al., 2023) address challenges like cross-frame deficiencies through pretrained image priors and dual-domain propagation. Specialized solutions like Raformer (Ji et al., 2025) target niche applications employing redundancy-aware attention to selectively process informative regions, enhancing both efficiency and accuracy. Although these models have made significant progress in video manipulation, they often fall short in downstream applications that demand high-fidelity photorealistic results, particularly in scenarios requiring precise content reconstruction. How to generate more precise and user-guided video edits is still a subject left for discussion.

Our proposed methodology addresses these challenges through an innovative inpainting pipeline that excels in preserving both temporal consistency and structural realism in dynamic settings. Through a real-time adaptive guidance framework, our approach automatically selects and updates guidance images throughout the video sequence, enabling precise restoration of fine-grained details without sacrificing temporal coherence. This evolving dual-guidance design significantly improves upon existing models, delivering enhanced realism and seamless inpainting performance in complex scenarios.

3 METHODS

3.1 Model Pipeline

In this section, we establish an automated, multi-image-guided, video-level diffusion-based inpainting pipeline. As illustrated in Figure 1, the pipeline consists of four interconnected modules.

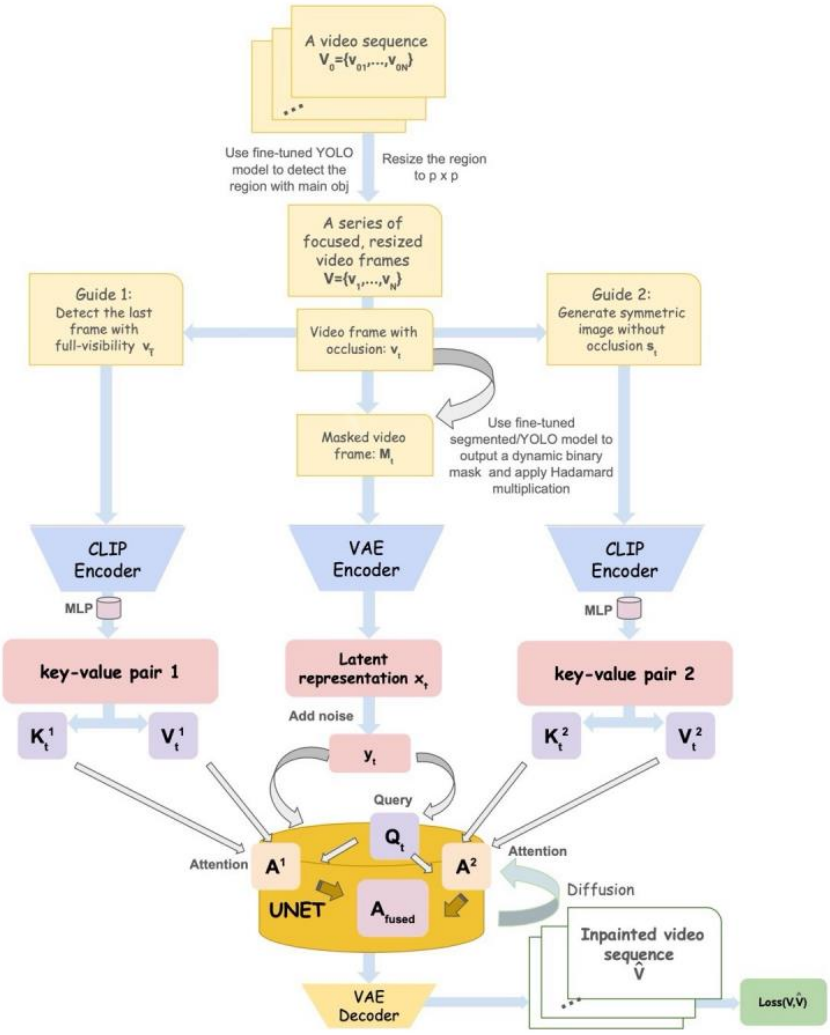

Figure 1. DiffMVR Model Pipeline.

The first module *Mod*$_1$, **Video Preprocessing,** detects and isolates the primary object in each frame using a fine-tuned YOLO model, ensuring that the inpainting process focuses accurately on regions of interest. This module prepares the input frames by resizing and aligning them for consistent processing. Additionally, we employ a fine-tuned YOLOv8-based model to detect bounding boxes or a segmentation model to identify irregular-shaped occlusions within the object.

The second module, *Mod*$_2$, **Visual Encoding**, independently encodes both the frame to be inpainted and its guidance images. The original video frame is processed through a VAE Encoder, which introduces noise to produce a latent

representation as input for the diffusion process. Concurrently, each guidance image, providing structural and temporal cues, is encoded by a CLIP Encoder to generate key-value pairs that facilitate subsequent attention mechanisms.

The third module, *Mod₃*, **Denoising with Fused Attention**, leverages spatial and temporal cues to guide the U-Net-based denoising process within the diffusion framework. By conditioning on the fused guidance information, this module enhances detail and continuity across frames, improving the overall output quality.

Finally, the fourth module, *Mod₄*, **Decoding and Restoration**, decodes the fully denoised frame representation back into pixel space using a VAE Decoder, producing the final inpainted frame. Each reconstructed frame is sequentially reassembled into the full video, yielding a temporally consistent inpainted video.

3.2 Problem Setting

We define the input video sequence as $V_0 = \{v_{0t}\}_{t=1}^{N}$, which is decomposed into sequential frames. Each frame v_{0t} undergoes processing to isolate the main object of interest, detected using a fine-tuned YOLOv8 model. The detected object in each frame is subsequently cropped and resized to a uniform resolution of $p \times p$, producing a refined video sequence $V = \{v_t\}_{t=1}^{N}$.

For inpainting facilitation, two guidance images are automatically generated for each frame v_t where occlusion is present: a symmetric image s_t and a past unobstructed frame $v_{\bar{t}}$. The symmetric image s_t is crafted by mirroring the unoccluded portion of v_t along an axis of symmetry, defined using Mediapipe (Kartynnik et al., 2019) for object landmark detection to precisely determine the symmetry line.

The past unobstructed frame $v_{\bar{t}}$ is sourced through a fine-tuned YOLOv8 model that scans previous frames in V for the most recently visible object, providing essential temporal guidance. Building on this, the construction of masked video frames M_t plays a critical role in isolating occluded regions for effective inpainting (YOLOv8 used). For complete details on the binary mask generation process and the construction of M_t from the input frames, please refer to Appendix *Binary Mask Methodology*.

We leverage both the VAE encoder and pre-trained CLIP image embeddings (Radford et al., 2021) to extract features for our inpainting pipeline. The masked video frame M_t is processed by the VAE encoder, transforming it into a spatial latent map x_t. Gaussian noise is then added to this map, producing a noisy latent y_t as preparation for iterative denoising within the U-Net.

Simultaneously, the guidance images, namely the symmetric reference $\{s_t\}_{t=2}^{N}$ and past unobstructed frames $\{v_{\bar{t}}\}_{\bar{t}=1}^{N-1}$, are encoded individually using the

CLIP encoders. Each guidance image is mapped from its original space to a *p*-dimensional feature vector, denoted as z_{s_t} and $z_{v_{\bar{t}}}$.

To ensure compatibility with the dimensions required for the diffusion module, each guidance embedding z_{s_t} and $z_{v_{\bar{t}}}$ is passed through a multi-layer perceptron(MLP), f_{mlp}: $\mathbb{R}^p \longrightarrow \mathbb{R}^{p'}$, which expands it to a p'-dimensional embedding:

$$\tilde{z}_{s_t} = f_{\text{mlp}}(z_{s_t}), \quad \tilde{z}_{v_{\bar{t}}} = f_{\text{mlp}}(z_{v_{\bar{t}}}).$$

The expanded embeddings generate key-value pairs (K_t^1, V_t^1) and (K_t^2, V_t^2) from each guidance image independently. These pairs contain spatial and temporal cues, which are then incorporated into the U-Net's denoising layers through cross-attention.

In Mod_3, at each U-Net layer, a query Q_t derived from the noisy latent y_t is used to compute attention scores A_t^1 and A_t^2, representing the relevance of each guidance source:

$$A_t^1 = \text{softmax}\left(\frac{Q_t K_t^{1\text{T}}}{\sqrt{D}}\right) V_t^1,$$

$$A_t^2 = \text{softmax}\left(\frac{Q_t K_t^{2\text{T}}}{\sqrt{D}}\right) V_t^2.$$

The final fused attention score A_{fused} combines A_t^1 and A_t^2 using weighted coefficients:

$$A_{\text{fused}}^t = \alpha_1 \cdot A_t^1 + \alpha_2 \cdot A_t^2.$$

We employ the dynamically computed A_{fused} score at each U-Net denoising layer, guiding the restoration process with high-level structural and temporal context. This innovation has proven its ability to overcome the continuity challenges in video inpainting.

During forward diffusion, noise is incrementally added to y_t, yielding

$$y_{t,T} = \sqrt{\bar{\alpha}_T} y_t + \sqrt{1 - \bar{\alpha}_T}\epsilon,$$

where $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise, and $\bar{\alpha}_T = \prod_{i=1}^{T} \alpha_i$ is the cumulative scaling factor for the noise component for $T = 1,2,\cdots$.

The U-Net's goal is to predict and remove the added noise at each timestep *T*. The diffusion loss is defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I),T}\left[\left\|\epsilon - \epsilon_\theta\left(y_{t,T}, T, A_{\text{fused}}^t\right)\right\|_2^2\right], \quad (1)$$

where $\epsilon_\theta\left(y_{t,T}, T, A_{\text{fused}}^t\right)$ represents the U-Net's prediction of the noise component conditioned on the input and fused attention at timestep *T*.

In the reverse diffusion process, the U-Net iteratively refines $y_{t,T}$ at each timestep *T*, aiming to reconstruct $y_{t,T-1}$:

$$y_{t,T-1} = \frac{1}{\sqrt{\alpha_T}}\left(y_{t,T} - \frac{1-\alpha_T}{\sqrt{1-\bar{\alpha}_T}}\epsilon_\theta\left(y_{t,T}, T, A_{\text{fused}}^t\right)\right) + \sigma_T z, \quad (2)$$

where $z \sim \mathcal{N}(0, I)$ and σ_T represents a noise scale factor at timestep T.
Upon completing the reverse diffusion process, the final denoised latent representation $\hat{y}_t$ is passed through the VAE decoder to reconstruct the inpainted frame:

$$\hat{v}_t = D(\hat{y}_t).$$

These reconstructed frames $\{\hat{v}_t\}_{t=1}^{N}$ are then sequentially reassembled to form the final inpainted video sequence $\hat{V} = \{\hat{v}_t\}_{t=1}^{N}$, ensuring temporal coherence and spatial fidelity throughout the sequence.

3.3 Loss Function

To achieve precise spatial inpainting while maintaining temporal coherence across video frames, we propose a combined loss function. This function is comprised of two components: the denoising loss, which focuses on spatial reconstruction, and the motion-consistency loss, which enforces smooth temporal transitions between frames in video sequences. The combined loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda \cdot \mathcal{L}_{\text{motion}}, \tag{3}$$

where λ is a weighting factor that balances the impact of temporal coherence against spatial accuracy.

Denoising Loss： Within the diffusion framework, the denoising term aims to iteratively restore each masked frame M_t by removing noise at each timestep T, as shown in (1).

Motion-Consistency Loss: We bring forward a motion-consistency loss to promote temporal coherence during denoising. At each timestep T, this loss measures the consistency between adjacent noisy frame representations

$$\mathcal{L}_{\text{motion}} = \frac{2}{N} \sum_{t=1}^{N-1} \left\| y_{t,T} - y_{t-1,T} \right\|_2^2, \tag{4}$$

where $y_{t,T}$ represents frame t at diffusion timestep T, and N is the total number of frames in the current video. This term encourages smooth transitions and consistent visual features across frames, reducing temporal artifacts introduced by independent frame processing.

The motion-consistency loss complements the denoising loss during the diffusion process, ensuring that the final video output achieves both high-quality spatial reconstruction and coherent temporal dynamics.

4 EXPERIMENTS

4.1 Implementation Details

Driven by the pressing need for an automated infant monitoring system that can reliably capture unobscured facial features in babies (Al-Yekreeti et al., 2024), we train and test our framework based on the iCOPEvid dataset (Brahnam et al., 2020).

The goal is to inpaint the occlusion by hand, and restore a clear, photorealistic baby face that provides healthcare professionals better decision support. This dataset contains 151 videos, each 20 seconds in duration, featuring 49 infants from various ethnic backgrounds. The videos capture the infants in a variety of states, including rest, friction, and pain. From these videos, we extracted 4,101 images of the same distribution, captured under a range of lighting intensities and health conditions (Kaduwela et al., 2024). These images are used to train the YOLO-based object detection model, both variants of the occlusion masking models, and the tuned text-to-image inpainting models, as well as to perform frame-level comparison tests.

DiffMVR builds upon the designs of stable-diffusion-v1-5 (Rombach et al., 2022), which introduce key architectural accommodations to tailor them for DiffMVR. For each video, we extract frames at 50 fps, yielding 1,000 frames per video across 120 videos in total. Next, we preprocess the frames following the designation we described in Mod_1. Finally, we partition the data into 70% for training (in total 63,000 frames), 10% for validation (in total 9,000 frames), and 20% for testing (in total 18,000 frames). We use the mean squared error (MSE) loss to reconstruct all pixels. Additionally, a motion loss is incorporated to improve both realism and temporal coherence in the video inpainting outputs.

Further details on implementations, including preprocessing of video frames, masking approaches, training strategies, and the formulation of our training loss, are available in Appendix *Extended Implementation Details*.

4.2 Baseline Models

To comprehensively compare our model's performance on static image inpainting and dynamic video inpainting, we perform both image-level and video-level tests. To evaluate video inpainting models on image-level tasks, we convert each test image into a 20-second video matching the input format these models require. Conversely, to apply image inpainting models to video-level tasks, we transform each video into a sequence of 1,000 frames, inpaint each frame individually, and then reassemble the frames into a complete video.

Image-level Comparison Models: We benchmark against LaMa (Suvorov et al., 2022) for image-guided inpainting, and both original and fine-tuned versions of Stabilityai (Rombach et al., 2022) and Runwayml (an open-source implementation that has been recently removed) models for text-guided inpainting. The fine-tuned variants (*Tuned-stabilityai* and *Tuned-runwayml*) are specifically adapted to the iCOPEvid dataset. In detail,

- *Tuned-stabilityai*: Fine-tuned from a general text-to-image stable diffusion model (Stabilityai) on 10 static images from the iCOPEvid dataset over 80 epochs using Dreambooth (Ruiz et al., 2023).

- *Tuned-runwayml*: Fine-tuned from a text-to-image stable diffusion model (Runwayml) on 1,575 frames from the iCOPEvid dataset over 200 epochs.

For the text-to-image inpainting models, we use the prompt “remove hands.” This prompt was chosen based on visualization experiments comparing different phrasings with the same meaning, where “remove hands.” produced the most accurate renderings. For the image-to-image model, we leverage the two masking models to generate both exact and roughly outlined masks.

Video-level Comparison Models: We compare DiffMVR with three advanced video inpainting models. The first model, End-to-End Flow-Guided Video Inpainting(E^{2}FGVI) (Li et al., 2022), integrates flow completion and latent feature propagation into a unified framework for inpainting. For E^{2}FGVI, we prepare the test data similarly to LaMa by using both masking models to generate masks for each video at 50 fps, resulting in 1,000 frames and 1,000 masks for each video sample.

The second model, Propainter by Zhou et al. (2023), is designed for efficient video restoration. It integrates dual-domain propagation with a mask-guided sparse video transformer that focuses on relevant regions, thereby reducing memory usage while delivering superior restoration performance. For each video sample, we prepare a single video (sampled at 50 fps) along with a corresponding set of frame-level masks (one mask per frame).

The third model by Wu et al. (2024), integrates multimodal large language models into a diffusion-based text-to-video inpainting framework. To employ this model, we adopt the prompt “remove the hands of the baby in the scene, replace with baby face,” which is selected based on both quantitative and qualitative evaluations. This model is purely text-guided and does not require any mask images.

4.3 Evaluation Metrics

We evaluate all models both qualitatively and quantitatively, focusing on both the independent images and continuous video frames. To demonstrate the robustness of our pipeline in capturing smooth transitions and restoring intricate details, we choose the following metrics: FID (Heusel et al., 2017), SSIM (Wang et al., 2004), TC (Lai et al., 2018), and FVD (Unterthiner et al., 2019). These metrics allow us to perform an all-rounded evaluation from three dimensions: structural similarity, the reality of restoration, and smoothness of the resulting video sequences. For details on the definition of these metrics, please refer to Appendix *Evaluation Metrics*.

5 QUANTITATIVE RESULTS

5.1 Frame-level

We leverage the 2,011 images extracted from the iCOPEvid dataset for the calculation of SSIM and FID scores. Additionally, we use 21 videos, each sampled at a frame extraction rate of 50 frames per second, for the calculation of the TC score.

Model	iCOPEvid Dataset – Segmented masks			iCOPEvid Dataset – Bounding boxes			HOF Dataset – Segmented masks		
	FID↓	SSIM↑	TC↓	FID↓	SSIM↑	TC↓	FID↓	SSIM↑	TC↓
DiffMVR	2.382	**0.899**	**0.395**	2.478	**0.864**	**0.396**	**5.412**	**0.786**	**0.428**
Stabilityai	3.066 ▲28.7%	0.686 ▼23.7%	0.424 ▲7.3%	3.230 ▲30.3%	0.706 ▼18.3%	0.430 ▲9.4%	6.118 ▲13.0%	0.751 ▼4.5%	0.430 ▲0.5%
Tuned-stabilityai	2.779 ▲16.7%	0.732 ▼18.6%	0.418 ▲5.8%	2.950 ▲19.0%	0.739 ▼14.5%	0.422 ▲6.6%	6.225 ▲15.0%	0.726 ▼7.6%	0.431 ▲0.7%
Runwayl	2.913 ▲22.3%	0.751 ▼16.5%	0.429 ▲8.6%	2.935 ▲18.4%	0.738 ▼14.6%	0.433 ▲10.2%	5.943 ▲9.8%	0.742 ▼5.6%	0.430 ▲0.5%
Tuned-runwayml	**2.366** ▼0.7%	0.763 ▼15.1%	0.424 ▲7.3%	**2.211** ▼10.8%	0.745 ▼13.8%	0.420 ▲6.1%	6.109 ▲12.9%	0.735 ▼6.5%	0.434 ▲1.4%
LaMa	2.940 ▲23.4%	0.712 ▼20.8%	0.455 ▲15.2%	3.105 ▲25.3%	0.670 ▼22.5%	0.457 ▲15.4%	7.025 ▲29.8%	0.731 ▼7.0%	0.456 ▲6.5%
E²FGVI	2.801 ▲17.6%	0.831 ▼7.6%	0.421 ▲6.6%	2.901 ▲17.1%	0.831 ▼3.8%	0.433 ▲10.2%	6.299 ▲16.4%	0.747 ▼5.0%	0.431 ▲0.7%
Propainter	2.773 ▲16.4%	0.850 ▼5.5%	0.397 ▲0.5%	2.853 ▲15.1%	0.836 ▼3.2%	0.399 ▲0.8%	6.345 ▲17.2%	0.762 ▼3.1%	0.429 ▲0.2%
LGVI	5.118 ▲114.9%	0.706 ▼21.4%	0.404 ▲2.3%	5.118 ▲106.5%	0.706 ▼18.3%	0.404 ▲2.0%	8.476 ▲56.6%	0.709 ▼9.8%	0.452 ▲5.6%
Gap	-0.68%	+17.82%	+6.84%	–12.08%	+15.97%	+5.71%	+11.41%	+6.94%	+1.38%
Gap between Masks	+3.87%	+4.05%	+0.25%	—	—	—	—	—	—

Note.

- Dash (—) = undefined value.
- ▲/▼ = relative increase/decrease in metric score compared to DiffMVR.
- The HOF Dataset (Ghanem et al., 2019) is used for proving the generality of DiffMVR.
- *Gap* = the extent by which DiffMVR outperforms (+) or is outperformed by (–) the second-best model (Tuned-runwayml).
- *Gap between Masks* = difference between segmented-mask vs. bounding-box results within DiffMVR.
- LGVI has identical mask/bbox scores because it is text-guided and does not use a mask image.

Table 1. Quantitative results comparing different models using FID, SSIM, and TC metrics on image-level for the iCOPEvid and HOF datasets.

As illustrated in Table 1, our model significantly outperforms the benchmark models in maintaining continuity between frames, as evidenced by the TC score, which surpasses the next best by 0.5%. Furthermore, achieving the highest overall metric scores across various datasets demonstrates our model's ability to capture detailed, realistic structures and ensures its robustness beyond our training dataset. Additionally, we observe an all-rounded better performance of segmented masks over bounding boxes in

Gap between Masks, which is expected since detailed images of parts of a human body come in irregular shapes, and thus bounding boxes mismatch.

By observing the numeric results in Table 1, we select the Tuned-runwayml as the second-best image-level model based on its consistent performance across the metrics. Additionally, relative comparisons (see the *Gap* column) show that DiffMVR outperforms Tuned-runwayml in most of the metrics, with the SSIM score improving by 17.8%, demonstrating the superior quality of DiffMVR's generation.

5.2 Video-level

As shown in Table 2, DiffMVR exhibits superior performance across both segmented masks and bounding boxes. Image-based models show notable

Model	iCOPEvid Dataset – Segmented masks				iCOPEvid Dataset – Bounding boxes			
	FID↓	SSIM↑	TC↓	FVD↓	FID↓	SSIM ↑	TC↓	FVD↓
DiffMVR	2.109	**0.905**	**0.338**	**47.88**	**2.102**	**0.881**	**0.339**	**50.40**
Stabilityai	2.406 ▲14.1%	0.738 ▼18.5%	0.421 ▲24.6%	73.94 ▲54.4%	2.497 ▲18.8%	0.736 ▼16.5%	0.427 ▲26.0%	74.39 ▲47.6%
Tuned-stabilityai	2.352 ▲11.5%	0.756 ▼16.5%	0.398 ▲17.8%	71.28 ▲48.9%	2.414 ▲14.8%	0.747 ▼15.2%	0.401 ▲18.3%	73.06 ▲45.0%
Runwayml	2.410 ▲14.3%	0.759 ▼16.1%	0.423 ▲25.1%	73.02 ▲52.5%	2.463 ▲17.2%	0.748 ▼15.1%	0.408 ▲20.3%	73.85 ▲46.5%
Tuned-runwayml	2.247 ▲6.5%	0.763 ▼15.7%	0.417 ▲23.4%	70.86 ▲48.0%	2.229 ▲6.0%	0.749 ▼15.0%	0.420 ▲23.9%	72.27 ▲43.4%
LaMa	2.933 ▲39.1%	0.720 ▼20.4%	0.454 ▲34.3%	77.95 ▲62.8%	3.195 ▲52.0%	0.695 ▼21.1%	0.455 ▲34.2%	78.12 ▲55.0%
E²FGVI	2.329 ▲10.4%	0.849 ▼6.2%	0.350 ▲3.6%	52.75 ▲10.2%	2.350 ▲11.8%	0.845 ▼4.1%	0.351 ▲3.5%	55.60 ▲10.3%
Propainter	**2.062** ▼2.2%	0.894 ▼1.2%	0.339 ▲0.3%	48.92 ▲2.2%	2.105 ▲0.1%	0.860 ▼2.4%	0.346 ▲2.1%	51.04 ▲1.3%
LGVI	4.545 ▲115.5%	0.717 ▼20.8%	0.346 ▲2.4%	56.20 ▲17.4%	4.545 ▲116.2%	0.717 ▼18.6%	0.346 ▲2.1%	56.20 ▲11.5%
Gap	−2.28%	+1.23%	+0.29%	+2.13%	+0.14%	+2.44%	+2.02%	+1.25%
Gap between Masks	−0.33%	+2.72%	+0.29%	+5.00%	—	—	—	—

Note.

• We have Propainter as the second-best model, for it has the majority of the second placement in metric values and good visualized results.

• *See the caption of Table 1 for other explanations.*

Table 2. Quantitative results comparing different models using FID, SSIM, TC, and FVD metrics on video-level for the iCOPEvid dataset.

degradation in object consistency and global alignment, affecting both frame-to-frame coherence and overall generation quality. Among the evaluated approaches, Propainter emerges as the second-best, particularly excelling in preserving spatial similarity. However, DiffMVR exceeds Propainter by 2.2%

on the FVD score, a comprehensive metric that evaluates video generation quality.

Moreover, DiffMVR consistently achieves the best SSIM, TC, and FVD scores by a substantial margin, highlighting its effectiveness in managing both the much larger iCOPEvid dataset and the smaller HOF dataset. To further illustrate DiffMVR's capability in generating naturalistic outputs, we present qualitative visualizations in the following section.

6 QUALITATIVE RESULTS

To demonstrate the efficacy of our approach, Figure 2 presents an in-distribution comparison between DiffMVR and the benchmark models across

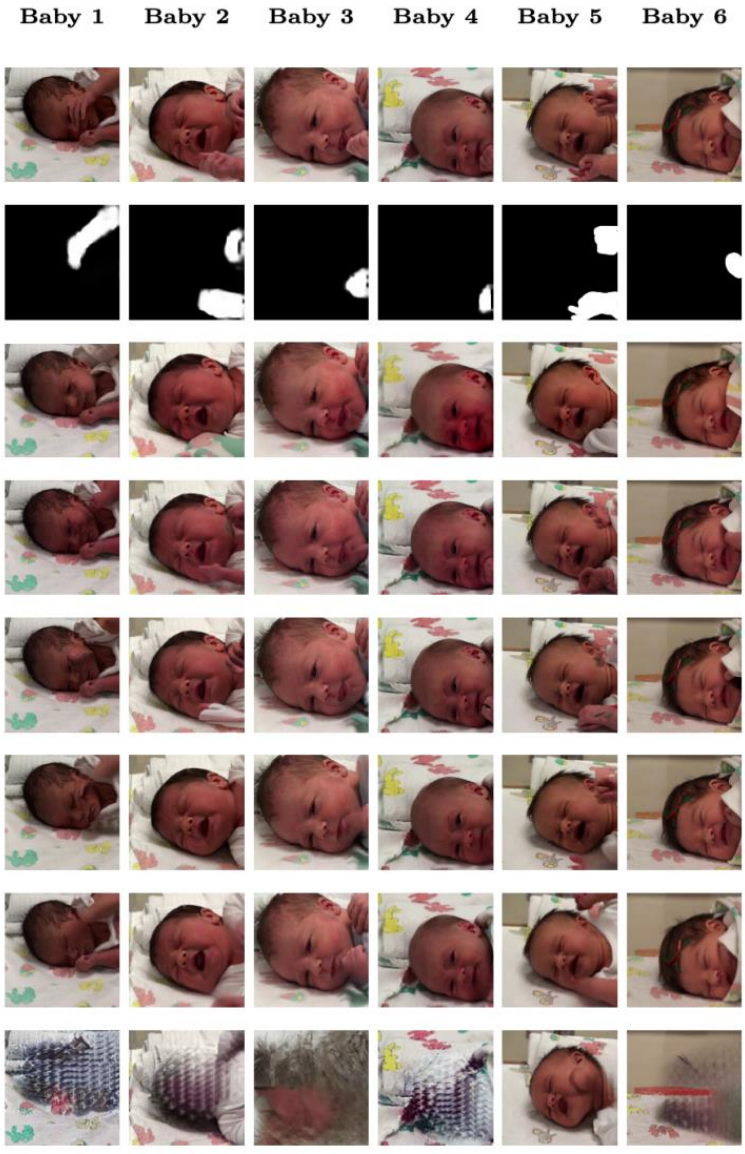

Figure 2. Qualitative comparison of DiffMVR with the benchmarked models on the iCOPEvid dataset.

videos with varying durations and masking complexities. DiffMVR uniquely satisfies multiple critical requirements: seamless blending of inpainted and original regions, complete obstruction removal, and accurate reconstruction of the baby's facial features without introducing incorrect anatomical elements. The model also maintains background integrity and ensures content consistency throughout the sequence. In contrast, other baseline models exhibit several shortcomings, such as distorted faces or backgrounds, incomplete removal of hands, restoration of incorrect hands (not belonging to the observed baby), low-resoluted restoration, and only partial removal of obstructions. This disadvantage is present even for the two second-best models *Tuned-runwayml* and *Propainter*.

Further displaying the robust performance of DiffMVR, Figure 3 highlights its ability in accurately restoring scenes on out-of-distribution images. We introduce the HandOverFace (HOF) dataset (Ghanem et al., 2019) as an additional test set. This dataset comprises of 302 images featuring various hand-over-face scenarios from a different distribution. Collected from publicly available sources, the HOF dataset represents diverse skin tones, motions, and age groups, enriching our evaluation with complex real-world cases. Looking at the test results on the HOF dataset, we have stronger evidence of DiffMVR's mightiness in capturing authentic details from a general viewpoint. Besides, for an illustrative example of the pipeline, please refer to Appendix *DiffMVR Pipeline*.

Figure 3. Occlusion removal and face restoration results on the HOF Dataset applying DiffMVR.

7 ABLATION STUDY

In the inpaint pipeline, we develop two key innovations: the dual-guidance module, which synthesizes fused embeddings from both short-term past and present frames to generate a new combined attention score, and the U-Net module, which designs and integrates a new motion-consistency loss term to guide the denoising process. In this section, we conduct a comprehensive ablation study to assess the effectiveness of having either or both modules in the pipeline.

7.1 Guidance Components Ablation

We contrast the performance of our model with variants that rely solely on a single-image guidance to illustrate the advantages of our multi-frame guidance module. This experiment specifically tests the impact of our innovative approach, which encodes guidance images independently and integrates them through a weighted cross-attention mechanism within the U-Net layers. Since from Table 2 segmented masks have better test results in the majority of aspects, we only compare results based on this masking type. As shown in Table 3, excluding either symmetric or prior guidance causes the inpaint result metrics to drop drastically, sometimes even worse than baseline models. Utilizing the current frame as guidance does not enhance the inpainting process, as evidenced by its subpar performance, ranking the last in comparison to benchmarks in both Table 2 and Table 3. By comparing DiffMVR against those restricted to a single type of guidance, we stress the necessity of the dual-image guidance design in our pipeline.

Model	Segmented masks			
	FID	SSIM	TC	FVD
Dual guide	2.11	0.91	0.34	47.88
Single guide (symmetric)	2.57 ▲21.8%	0.75 ▼17.6%	0.42 ▲23.5%	59.51 ▲24.3%
Single guide (past frame)	2.36 ▲11.8%	0.77 ▼15.4%	0.38 ▲11.8%	60.80 ▲27.0%
Single guide (present frame)	2.95 ▲39.8%	0.72 ▼20.9%	0.45 ▲32.4%	72.79 ▲52.0%

Note. This table highlights the design of multi-guidance achieves the best performance. The motion loss is included throughout this comparison test. The ▲/▼ indicates a relative increase/decrease in metric score compared to dual guide (*DiffMVR)*.

Table 3. Guidance component ablation test on the iCOPEvid dataset.

7.2 Loss Component Ablation

Building upon the findings from *Guidance Components Ablation*, this ablation study further investigates the cumulative impact of integrating the additional motion loss component into our pipeline. To systematically assess the impact of each component, we conduct experiments under several configurations. Using a single past frame as guidance and using merely denoise loss for training is the baseline setting. We gradually add the designs in: i) baseline + dual-guide, ii) baseline + motion loss, and iii) baseline + dual-guide + motion loss, which is our model, DiffMVR.

We present the results in Table 4. As expected, adding the motion-consistency loss leads to a lower TC score and higher FVD compared to the baseline, even when a single image is used as guidance.

Adding motion as a portion of loss enhances temporal smoothness and contributes to a more natural video frame restoration, revealed by the scope of changes in the row of *Gap*. Besides, the comparison between *baseline* and *DiffMVR* shows that incorporating both the motion loss and using dual guidance notably improves the model's performance by 15.2% on average. These results confirm that each component of our approach works cooperatively, with the full integration of all elements necessary for an optimal performance.

Configuration	Segmented Masks			
	FID	SSIM	TC	FVD
baseline	2.46	0.68	0.41	65.92
baseline + dual	2.28 ▼7.3%	0.74 ▲8.8%	0.37 ▼9.8%	61.57 ▼6.6%
baseline + motion	2.36 ▼4.1%	0.77 ▲13.2%	0.38 ▼7.3%	60.80 ▼7.8%
DiffMVR: baseline + dual + motion	2.11 ▼14.2%	0.91 ▲33.8%	0.34 ▼17.1%	47.88 ▼27.4%
Gap (%)	7.46	22.97	8.11	22.23

Note. This table exemplifies the impact of gradually adding motion-consistency loss to different guidance configurations. The results highlight the combined effect of our innovations in enhancing video inpainting performance. ▲/▼ indicates a relative increase/decrease in metric score compared to baseline. *Gap* refers to the extent by which *DiffMVR: baseline + dual + motion* outperforms *baseline + dual*.

Table 4. Loss component ablation test on the iCOPvid dataset.

8 CONCLUSION

In this study, we introduced multi-image guidance for inpainting, a more intuitive alternative where images speak louder than words. Nonetheless, opportunities for further exploration exist, particularly in optimizing the weighting factor λ to better match user preferences and application requirements. We hope DiffMVR will drive innovations in video processing and stimulate further advances in downstream inpainting tasks.

FUNDING

This research is funded by the U.S. National Science Foundation (NSF) under Grant #2205472.

ACKNOWLEDGMENT

The iCOPEvid (Infant COPE Database) is copyrighted by Dr. Sheryl Brahnam and is used with her permission.

REFERENCES

Al-Tekreeti, Z., Moreno-Cuesta, J., Garcia MIM., & Rodrigues, MA. (2024). AI-based visual early warning system. *Informatics*, 11(3), 59. https://doi.org/10.3390/informatics11030059.

Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18187-18197). https://doi.org/10.1109/CVPR52688.2022.01767.

Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18187-18197. https://doi.org/10.1109/CVPR52688.2022.01767.

Brahnam, S., Nanni, L., McMurtrey, S., Lumini, A., Brattin, R., Slack, M., & Tonya, B. (2020). Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from gaussian of local descriptors. *Appllied Computing and Informatics*, 19, 122-143. https://doi.org/10.1016/j.aci.2019.05.003.

Ghanem, S., Imran, A., & Athitsos, V. (2019). Analysis of hand segmentation on challenging hand over face scenario. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 236-242). https://doi.org/10.1145/3316782.3321534.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144. https://doi.org/10.1145/3422622.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the International Conference on Neural Information Processing Systems* (pp. 6629–6640).

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840-6851.

Hsu, CC., Lee, CM., & Chou, YS. (2024). DRCT: saving image super-resolution away from information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6133-6142). https://doi.org/10.1109/CVPRW63382.2024.00618.

Ji, Z., Su, Y., Zhang, Y., Hou, J., Pang, Y., & Han, J. (2025). Raformer: redundancy-aware transformer for video wire inpainting. *IEEE Transactions on Image Processing*, 34, 1795-1809. https://doi.org/10.1109/TIP.2025.3550038.

Kaduwela, NA., Horner, S., Dadar, P., & Manworren, R. (2024). Application of a human-centered design for embedded machine learning model to develop data labeling software with nurses: human-to-artificial intelligence (H2AI). *International Journal of Medical Informatics*, 183. https://doi.org/10.1016/j.ijmedinf.2023.105337.

Kartynnik, Y., Ablavatski, A., Grishchenko, I., & Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile GPUs. In *Proceedings of the IEEE/CVF Workshop on Computer Vision for Augmented and Virtual Reality*. https://doi.org/10.48550/arXiv.1907.06724.

Kim, G., Kwon, T., & Ye, J.C. (2022). DiffusionCLIP: text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2426-2435). https://doi.org/10.1109/CVPR52688.2022.00246.

Lai, WS., Huang, JB., Wang, O., Shechtman, E., Yumer, E., & Yang, MH. (2018). Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision* (pp. 179-195). https://doi.org/10.1007/978-3-030-01267-0_11.

Lee, M., Cho, S., Shin, C., Lee, J., Yang, S., & Lee, S. (2025). Video diffusion models are strong video inpainter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4), 4526-4533. https://doi.org/10.1609/aaai.v39i4.32477.

Li, Z., Lu, CZ., Qin, J., Guo, C., & Cheng, M. (2022). 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). https://doi.org/10.1109/CVPR52688.2022.01704.
Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1833-1844). https://doi.org/10.1109/ICCVW54120.2021.00210.
Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: towards photorealistic Image generation and editing with text-guided diffusion models. In *Proceedings of the 35th International Conference on Machine Learning*, 162, 16784-16804. https://doi.org/10.48550/arXiv.2112.10741.
Ouyang, H., Wang, T., & Chen, Q. (2021). Internal video inpainting by implicit long-range propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp.14559-14568). https://doi.org/10.1109/ICCV48922.2021.01431.
Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. (2016). Context encoders: feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2536-44). https://doi.org/10.1109/CVPR.2016.278.
Radford, A., Kim, JW., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Jack, C., Kruegar, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748-63. https://doi.org/10.48550/arXiv.2103.00020.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684-10695). https://doi.org/10.1109/CVPR52688.2022.01042.
Ruiz N, Li Y, Jampani V, Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22500-22510). https://doi.org/10.1109/CVPR52729.2023.02155.
Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., & Norouzi, M. (2022). Palette: image-to-image diffusion models. In *SIGGRAPH 22: Special Interest Group on Computer Graphics and Interactive Techniques*, 15, 1-10. https://doi.org/10.1145/3528233.3530757.
Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. *International Conference on Learning Representations*.
Susladkar, O., SenGupta, J., Sehgal, C., Mittal, S., & Singhal, R. (2025). MotionAura: generating high-quality and motion consistent videos using discrete diffusion. In *Proceedings of the International Conference on Learning Representations*.
Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A. & Silvestrov, A. (2022). Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2149-2159). https://doi.org/10.1109/WACV51458.2022.00323.
Unterthiner, T., Steenkiste SV, Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2019). Towards accurate generative models of video: a new metric & challenges. In *Proceedings of the International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1812.01717.
Wang, Z., Bovik, AC., Sheikh, HR., & Simoncelli, EP. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612. https://doi.org/10.1109/TIP.2003.819861.
Wu, JZ., Ge, Y., Wang, X., Lei, SW., Gu, Y., & Shi, Y. (2023). 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). https://doi.org/10.1109/ICCV51070.2023.00701.

Wu, J., Li, X., Si, C., Zhou, S., Yang, J., & Zhang, J. (2024). 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). https://doi.org/10.1109/CVPR52733.2024.01188.

Xu, Z., Zhang, X., Chen, W., Yao, M., Liu, J., Xu, T., & Wang, Z. (2023). A review of image inpainting methods based on deep learning. *Applied Sciences*, 13(20), 11189. https://doi.org/10.3390/app132011189.

Yang, L., Zhang, Z., Yu, Z., Liu, J., Xu, M., Ermon, S., & Cui, B. (2024). Cross-modal contextualized diffusion models for text-guided visual generation and editing. In *Proceedings of the International Conference on Learning Representations (Poster)*. https://openreview.net/forum?id=nFMS6wF2xq.

Zhang, R., Isola, P., & Efros, A.A. (2016). Colorful image colorization. In *Proceedings of the European Conference on Computer Vision* (pp. 649-666). https://doi.org/10.1007/978-3-319-46487-9_40.

Zhang, Z., Wu, B., Wang, X., Luo, Y., Zhang, L., & Zhao, Y. (2024). AVID: any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7162-7172). https://doi.org/10.1109/CVPR52733.2024.00684.

Zhou, S., Li, C., Chan, KCK., & Loy, CC. (2023). ProPainter: improving propagation and transformer for video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 10443-10452). https://doi.org/10.1109/ICCV51070.2023.00961.

Supplementary Material for "DiffMVR: An Intelligent Diffusion-based Multi-Guidance System for Real-time Video Restoration"

Zheyan Zhang, Diego Klabjan, Renee CB Manworren

May 15, 2025

Contents

A. Extended Literature Review

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and their variants (Brock et al., 2018) have transformed image editing by generating high-quality, realistic images and enabling unsupervised learning. Advanced GAN-based techniques, such as the approach by Han et al. (2018), which incorporates contextual attention and a dual-discriminator framework, and the Aggregated Contextual-Transformation GAN (Zeng et al., 2023), which combines transformations from multiple receptive fields to enhance texture synthesis, further improve restoration quality. Following the emergence of GANs, patch-based methods (Yu et al., 2018) were introduced to synthesize textures from undamaged regions, although they struggle with larger missing areas and maintaining global coherence in complex scenes.

Building on these earlier inpainting techniques, recent research has shifted focus toward diffusion models as a robust alternative for image restoration. Pioneering approaches like Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) and their variants (e.g., Denoising Diffusion Implicit Models, DDIM (Song et al., 2021)) use iterative denoising to achieve high-quality restoration. Complementing these techniques, Vector Quantized Variational AutoEncoder (VQ-VAE) (Razavi et al., 2019) pushes the boundaries of high-resolution image synthesis by learning quantized embeddings, though currently constrained to single-frame generation. Meanwhile, partial convolution inpainting (Liu et al., 2018) addresses irregular masks by conditioning filters on valid pixels alone, reducing artifacts in static tasks.

Diffusion-based inpainting has demonstrated remarkable versatility across diverse fields, showcasing its potential beyond traditional image restoration techniques. In medical imaging, these models aid in anomaly detection by restoring diseased regions for comparative analysis (Wolleb et al., 2022). In autonomous driving, they reconstruct occluded road signs (Liu et al., 2025), while in advertising, they create immersive VR scenes for product promotion (Asija et al., 2024). They also

facilitate privacy preservation by removing sensitive visual details (e.g., faces or personal data), and in healthcare, they enable real-time facial action monitoring for pain assessment (Herr et al., 2024).

However, a critical limitation remains evident across most current inpainting methodologies. Despite their impressive capabilities in realistic object replacement and seamless frame-level integration, these approaches predominantly focus on static images. This frame-centric perspective creates a fundamental constraint: these methods often struggle to capture the dynamic changes and temporal nuances inherent in continuous video streams. The resultant gap highlights an urgent need for advanced video-level diffusion models capable of simultaneously integrating spatial and temporal contextual cues, promising a more holistic approach to image and video restoration.

Early approaches, such as those developed by Li et al. (2022), employ flow-based techniques and deformable convolutions to propagate features and enforce frame continuity, yet their reliance on intermediate flow estimation can introduce cumulative errors. DNN-based inpainting models, including the Copy-and-paste network (Lee et al., 2019) and the context-aggregated network (Li et al., 2020), address context restoration by aggregating reference frames.

B. Motivation

Our approach to video inpainting, DiffMVR, develops a sophisticated reconstruction mechanism that transcends traditional frame-based restoration techniques. At the core of our methodology lies an intricate dual-guidance image generation strategy designed to capture both spatial and temporal nuances of occluded video content. For each masked frame, we simultaneously generate two critical guidance images: a symmetric image and a past unobstructed frame. The symmetric image emerges through a precise mirroring process, reflecting the visible half of the frame along its central axis, which provides structural insights into the frame's underlying composition. Concurrently, we deploy a fine-tuned YOLOv8 model to identify and extract the most recent fully visible instance of the occluded object from preceding frames, thereby establishing a temporal reference point for reconstruction.

The reconstruction process unfolds through a meticulously designed multi-stage approach that integrates advanced machine learning techniques. Both guidance images undergo processing by separate CLIP models, enabling the extraction of sophisticated key-value feature pairs that capture semantic and structural information. The current masked frame is encoded into a latent space utilizing a Variational Autoencoder (VAE), with carefully introduced random noise serving as a query mechanism. This query dynamically interacts with the extracted key-value pairs, generating dual attention scores that are subsequently weighted and strategically fused. The U-Net architecture then leverages this combined attention, alongside standard diffusion inputs, to iteratively denoise and recover a pristine latent vector, which the VAE ultimately decodes into a restored frame representation.

Recognizing the critical importance of maintaining temporal coherence, we introduce an innovative motion loss term that fundamentally transforms the video reconstruction process. Unlike traditional image-level denoising methods that treat each frame in isolation, our non-separable frame loss function creates intricate linkages between adjacent frame representations. This approach ensures a continuous and seamless video reconstruction that preserves the fluidity of actions and authentically represents the scene's dynamic characteristics. By tightly connecting each frame to its predecessors and successors, we create a unified video sequence that seamlessly integrates spatial details and temporal dynamics, offering unprecedented precision in object restoration and scene reconstruction.

C. Binary Mask Methodology

DiffMVR leverages binary masking approach that captures the intricacies of frame occlusion. We define a binary mask $m_{t,i}$ for each video frame v_t, where each pixel is categorically classified:

$$m_{t,i} = \begin{cases} 1, & \text{if pixel } i \text{ is part of the occlusion,} \\ 0, & \text{otherwise.} \end{cases}$$

In our model, we employ two different mask-generation techniques tailored for continuous video frames. The first mask generation model is based on a YOLOv8 structure and produces bounding box masks. The second model adapts a segmentation-based approach (Camporese et al., 2021) and provides irregularly contoured masks. These are parts of preprocessing in Mod_1. We train and test the pipeline on both types, and both results are presented in the experimental section.

These mask generation techniques form a critical preprocessing component of our first module (Mod_1). Once the binary masks $m_{t,i}$ are generated, we construct masked video frames through a precise mathematical operation:

$$M_t = m_{t,i} \odot v_{t,i}, \forall i \in \text{pixels}, t \in \{2, \ldots, N\},$$

where $\odot$ denotes the Hadamard product, preserving only the regions indicated by the mask in each frame v_t.

The culmination of this preprocessing stage results in a set of processed frames: M_t (masked frames), s_t (supplementary information), and $v_{\bar{t}}$ (reference frames), spanning the video sequence. At the end of Mod_1, the processed frames M_t, s_t, and $v_{\bar{t}}$, for $t \in \{2, \ldots, N\}$, $\bar{t} \in \{1, \ldots, N-1\}$, are passed to the next module, which encodes spatial and temporal cues into compact representations.

D. Extended Implementation Details

Our experimental framework begins with preprocessing of both the video frames and images, first centered on the facial region and then resized to 512×512 pixels. This operation is performed by the fine-tuned YOLOv8-based model, which demonstrates a 100.0% accuracy in detecting the main object, in our case the infant's face. Eventually, all processed data are standardized to this resolution for training, evaluation, and benchmark comparisons.

To train and evaluate our model's robustness under different occlusion scenarios, we implement two distinct masking approaches. The first employs the fine-tuned YOLOv8n model (Terven et al., 2023), trained on 96 annotated images from 4 babies in the ICOPEvid dataset and tested on 25 images from 2 babies. This model achieves 97.5% masking accuracy and an average IoU of 0.979, generating rectangular masks for occlusions. The second method leverages a fine-tuned custom segmentation model (Camporese et al., 2021) trained on 215 rigorously labeled images from 5 babies, reaching 96.4% accuracy and producing irregular-shaped masks with an average IoU of 0.930, better mimicking real-world occlusions.

We train with Adam optimizer, setting the learning rate 10^{-5}, with a batch size of 8, then trained for $420,000$ iterations. We implement our method using the PyTorch (version $v2.2.2$) framework. We use the following hyperparameter names, consistent with PyTorch conventions:

```
model:
  params:
    linear_start: 0.0008
    linear_end: 0.01450
    num_timesteps_cond: 1
    timesteps: 1000
    conditioning_key: fused crossattn

  trainer:
    type: "Adam"
    base_learning_rate: e-5
    warm_up_steps: 1000
    batch_size: 8
    log_freq: 500
    val_log_freq: 2e3
    iterations: 420e3
```

E. Evaluation Metrics

We perform the image-level evaluation using the static images from the ICOPEvid dataset, where the ground truth is the original image, while the prediction is the inpainted image.

We assess the spatial similarity with the Structural Similarity Index Measure (SSIM), which measures the similarity between the ground truth frames $V = \{v_t\}_{t=1}^{N}$ and the corresponding reconstructed frames $\hat{V} = \{\hat{v}_t\}_{t=1}^{N}$, by comparing luminance, contrast, and structural properties of the pixels using an 11×11 Gaussian window W. The final score is obtained by averaging the SSIM values across all N frames, calculated as

$$\text{Mean SSIM} = \frac{1}{N} \sum_{t=1}^{N} \text{SSIM}(v_t, \hat{v}_t),$$

where

$$\text{SSIM}(v_t, \hat{v}_t) = \frac{(2\mu_{v_t}\mu_{\hat{v}_t} + C_1)(2\sigma_{v_t\hat{v}_t} + C_2)}{(\mu_{v_t}^2 + \mu_{\hat{v}_t}^2 + C_1)(\sigma_{v_t}^2 + \sigma_{\hat{v}_t}^2 + C_2)}, \tag{E.1}$$

- μ_{v_t} and $\mu_{\hat{v}_t}$: the mean pixel values over window W,
- $\sigma^2_{v_t}$ and $\sigma^2_{\hat{v}_t}$: the variances of each frame in the window,
- $\sigma_{v_t\hat{v}_t}$: covariance between v_t and $\hat{v}_t$,
- $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$: constants to stabilize division with weak denominator,
- L: dynamic range of pixel values,
- $k_1 = 0.01$ and $k_2 = 0.03$ by default.

To evaluate the realism of generated frames, we calculate the Fréchet Inception Distance (FID), comparing the distributions of real images $V = \bigcup_{i=1}^{N} v_i$ and generated images $\hat{V} = \bigcup_{i=1}^{N} \hat{v}_i$ in the feature space of an Inception network. FID is computed as

$$\text{FID}(V, \hat{V}) = \|\mu - \hat{\mu}\|_2^2 + \text{Tr}\left(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{\frac{1}{2}}\right), \tag{E.2}$$

where

- μ and $\hat{\mu}$ are the mean feature vectors of the true and inpainted image,
- Σ and $\hat{\Sigma}$ are the covariance matrices of the true and inpainted image distributions.

At last, we measure the temporal coherence. For each video, the Temporal Consistency (TC) score is calculated based on the normalized squared $L2$ norm difference between consecutive frames. The overall TC score is obtained by averaging the TC scores across all videos. For each neighboring frame pair, $\hat{v}_t$ and $\hat{v}_{t+1}$ within the same video, we first flatten each frame into a vector, and then compute the TC score for each video as

$$\text{TC} = \frac{1}{N-1} \sum_{t=1}^{N-1} \left(\frac{\|d_{t+1} - d_t\|_2^2}{512^2 \times 255^2} \right), \tag{E.3}$$

where

- N is the total number of frames in the video,
- d_t represents the vectorized form of frame $\hat{v}_t$.

We extend the use of SSIM, FID, and TC metrics to evaluate the Infant video dataset. These metrics are computed for each video, and their scores are averaged across all videos. The resulting average scores are denoted as $\bar{\text{SSIM}}$, $\bar{\text{FID}}$, and $\bar{\text{TC}}$ for simplicity and clarity in reporting.

Additionally, we introduce the Frechet Video Distance (FVD) metric for a pervasive assessment at the video level. FVD builds on the concept of the FID score by incorporating video-based features extracted using a pre-trained Inflated 3D Convolutional Neural Network (I3D). This adaptation allows FVD to capture both video content's spatial and temporal dynamics. The calculation of FVD follows the formula presented in E.2, substituting frame-level features with video-level features extracted by I3D as follows

$$\text{FVD}(\mathbf{V}_0, \hat{\mathbf{V}}_0) = \left\|\boldsymbol{\mu}_{\mathbf{V}_0} - \boldsymbol{\mu}_{\hat{\mathbf{V}}_0}\right\|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{V}_0} + \boldsymbol{\Sigma}_{\hat{\mathbf{V}}_0} - 2(\boldsymbol{\Sigma}_{\mathbf{V}_0}\boldsymbol{\Sigma}_{\hat{\mathbf{V}}_0})^{\frac{1}{2}}\right), \tag{E.4}$$

where

- $\mathbf{V}_0$ and $\hat{\mathbf{V}}_0$: sets of videos,
- $\boldsymbol{\mu}_{\mathbf{V}_0}$ and $\boldsymbol{\mu}_{\hat{\mathbf{V}}_0}$: mean feature vectors extracted from the videos,
- $\boldsymbol{\Sigma}_{\mathbf{V}_0}$ and $\boldsymbol{\Sigma}_{\hat{\mathbf{V}}_0}$: covariance matrices of the video feature distributions.

F. DiffMVR Pipeline

Figure F.1 presents an example of the DiffMVR pipeline. It showcases an input video, the segmented masks, and the guidance images (1 detected through the same video source and 2 generated from the input frame accordingly), as depicted in rows 1, 2, 3, and 4. Row 5 highlights the final inpainted results, where occluded hand regions are removed and replaced with a realistically restored face. Although this focus on face restoration is particularly challenging, as distortions and artifacts tend to be more common in facial inpainting than in object-based tasks, DiffMVR managed to accurately detect and remove occluded hand regions, filling them with realistic facial details.

Besides seamlessly completing occluded regions with coherent, clear content, DiffMVR adeptly handles challenging conditions, such as low lighting, and excels at editing varied object textures and colors, and is demonstrated in Figure F.2, revealing its flexibility in diverse inpainting scenarios.

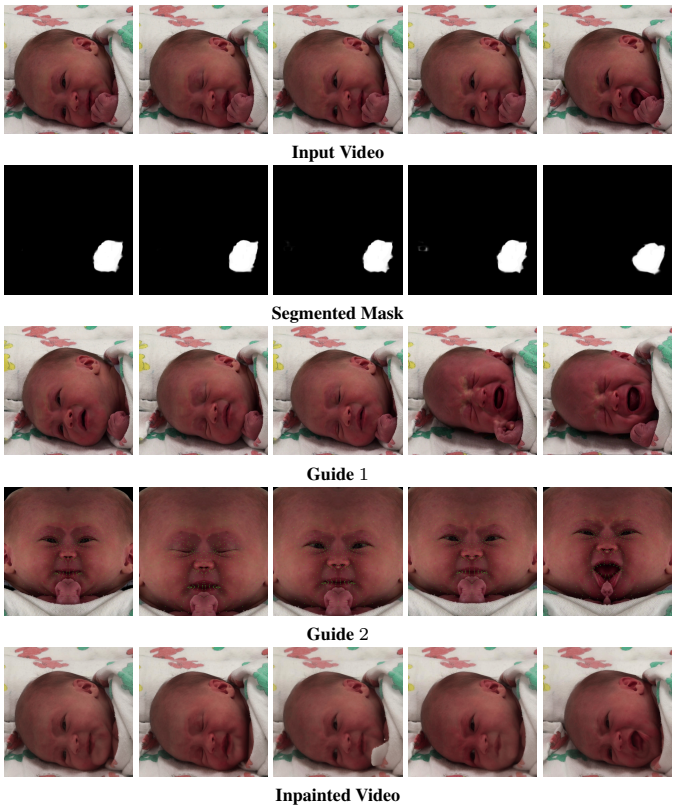

Figure F.1. DiffMVR's illustrative example on *S024_Pain_20s.mp4*. The first row input frames are from iCOPEvid (Copyrighted by Dr Sheryl Brahnam. Used and reprinted with permission).

Figure F.2. DiffMVR model's flexibility in both image condition and inpaint function. The left column shows the input video frame, and the right shows the inpaint result under proper guidance images design.

References

[1] Asija, S., Du, E., Nguyen, N., Zollmann, S., & Ventura, J. (2024). 3D pano inpainting: building a vr environment from a single input panorama. *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops* (pp. 1019–1020). https://doi.org/10.1109/VRW62533.2024.00306

[2] Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. In *Proceedings of the International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:52889459

[3] Camporese, G. (2021). Hands segmentation is all you need. https://github.com/guglielmocamporese/hands-segmentation-pytorch.

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *Communications of the ACM*, 63(11), 2672–2680. https://doi.org/10.1145/3422622.

[5] Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, LS. (2018). VITON: an image-based virtual try-on network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7543–7552). https://doi.org/10.1109/CVPR.2018.00787.

[6] Herr, K., Anderson, AR., Arbour, C., Coyne, PJ., Ely, E., Gelinas, C., & Manworren, R. (2024). Pain assessment in the patient unable to self-report: clinical practice recommendations in support of the ASPMN 2024 position statement. *Pain Management Nursing*, 25(6), 551–568. https://doi.org/10.1016/j.pmn.2024.09.010.

[7] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.

[8] Lee, S., Oh, S. W., Won, D., & Kim, SJ. (2019). Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4412–4420). https://doi.org/10.1109/ICCV.2019.00451.

[9] Li, A., Zhao, S., Ma, X., Gong, M., Qi, J., Zhang, R., Tao, D., & Kotagiri, R. (2020). Short-term and long-term context aggregation network for video inpainting. In *Proceedings of the European Computer Vision Association* (pp. 728–743). https://doi.org/10.1007/978-3-030-58548-8_42.

[10] Li, Z., Lu, CZ., Qin, J., Guo, C., & Cheng, M. (2022). 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). https://doi.org/10.1109/CVPR52688.2022.01704.

[11] Liu, G., Reda, FA., Shih, KJ., Wang, TC., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision* (pp. 89–105). https://doi.org/10.1007/978-3-030-01252-6_6.

[12] Liu, J., Hang, P., Zhao, X., Wang, J., & Sun, J. (2025). DDM-lag: a diffusion-based decision-making model for autonomous vehicles with lagrangian safety enhancement. *IEEE Transactions on Artificial Intelligence*, 6(3), 780–791. https://doi.org/10.1109/TAI.2024.3497918.

[13] Razavi, A., Oord, VVA., & Vinyals, O. (2019). Generating diverse high-fidelity images with VQ-VAE-2. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 14866–14876). https://dl.acm.org/doi/10.5555/3454287.3455618.

[14] Song, J., Meng, C. & Ermon, S. (2021). Denoising diffusion implicit models. *International Conference on Learning Representations*.

[15] Terven, J., Córdova-Esparza, DM., & Romero-González, JA. (2023). A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4), 1680–1716. https://doi.org/10.3390/make5040083.

[16] Wolleb, J., Bieder, F., Sandkühler, R., & Cattin, PC. (2022). Diffusion models for medical anomaly detection. *International Conference on Medical Image Computing and Computer Assisted Intervention*, 13438, 35–45. https://doi.org/10.1007/978-3-031-16452-1_4.

[17] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, TS. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5505–5514). https://doi.org/10.1109/CVPR.2018.00577.

[18] Zeng, Y., Fu, J., Chao, H., & Guo, B. (2023). Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 29, 3266–3280. https://doi.org/10.1109/TVCG.2022.3156949.