Video Diffusion Models are Strong Video Inpainter

Minhyeok Lee, Suhwan Cho, Chajin Shin, Jungho Lee, Sunghun Yang, Sangyoun Lee

Yonsei University
{hydragon516,chosuhwan,chajin,2015142131,sunghun98,syleee}@yonsei.ac.kr

Abstract

Propagation-based video inpainting using optical flow at the pixel or feature level has recently garnered significant attention. However, it has limitations such as the inaccuracy of optical flow prediction and the propagation of noise over time. These issues result in non-uniform noise and time consistency problems throughout the video, which are particularly pronounced when the removed area is large and involves substantial movement. To address these issues, we propose a novel First Frame Filling Video Diffusion Inpainting model (FFF-VDI). We design FFF-VDI inspired by the capabilities of pre-trained image-to-video diffusion models that can transform the first frame image into a highly natural video. To apply this to the video inpainting task, we propagate the noise latent information of future frames to fill the masked areas of the first frame's noise latent code. Next, we fine-tune the pre-trained image-to-video diffusion model to generate the inpainted video. The proposed model addresses the limitations of existing methods that rely on optical flow quality, producing much more natural and temporally consistent videos. This proposed approach is the first to effectively integrate image-to-video diffusion models into video inpainting tasks. Through various comparative experiments, we demonstrate that the proposed model can robustly handle diverse inpainting types with high quality.

Introduction

Video inpainting (VI) aims to fill in missing regions of a video with visually consistent content while ensuring both spatial and temporal consistency. Unlike image inpainting, video inpainting faces the significant challenge of establishing accurate correspondences with distant frames to obtain missing pixel information and maintaining time consistency between video frames to generate natural videos. To address these issues, recent methods (Gao et al. 2020; Zhang, Fu, and Liu 2022b,a; Zhou et al. 2023) based on information propagation through optical flow have gained significant attention. These models use inpainted optical flow information to propagate pixel or feature information to the masked areas of each frame. While this approach can generate relatively accurate images by propagating actual pixel information to the missing regions, it has several significant drawbacks. Firstly, the quality of the predicted video is highly de-

pendent on the accuracy of the completed optical flow, since pixel information is propagated based on the optical flow between all adjacent frames. Incorrect predictions can accumulate error noise during the frame propagation process, reducing the overall quality of the video frames. Secondly, in the case of object removal, propagation-based models require relatively accurate target object masks. These methods are highly dependent on the amount of visual cues from the unmasked areas of the reference frames to fill the masked regions of the target frame. Therefore, to ensure consistent performance, the mask size must be minimized, which necessitates precise target object masks. However, this leads to increased manual frame-by-frame masking costs in real-world applications.

Recently, a few works (Zhang et al. 2024; Zi et al. 2024; Gu et al. 2023) integrate the powerful generative capabilities of image or video diffusion models into video inpainting tasks. These methods are expected to be more robust to mask size and generate perceptually improved videos compared to propagation-based methods. However, despite the advancements in diffusion models, there are two main reasons that make it challenging to apply them to video inpainting. First, diffusion models do not consider the objects behind the moving mask areas. As the mask's position shifts over time, the actual pixel information in the previously masked areas is revealed. If the past frames generated by the diffusion model differ from the real context, this leads to videos with awkward consistency over time. This becomes even more unnatural when new objects hidden behind the mask appear. Second, when using pre-trained video diffusion models, object hallucination often occurs, where unwanted objects are generated in the mask area. This phenomenon is due to the diffusion model's powerful video generation capabilities, and to prevent it, text guidance is required for the diffusion model.

To address these issues, we propose a new First Frame Filling Video Diffusion Inpainting model, named FFF-VDI. We design FFF-VDI, inspired by the capabilities of pre-trained image-to-video diffusion models to generate complete videos based on the first frame image. Our model has the following advantages: First, the proposed model extracts the latent code of each frame using a VAE encoder to create the masked noise latent code. Next, we use optical flow to propagate the noise latent codes of future frames to fill
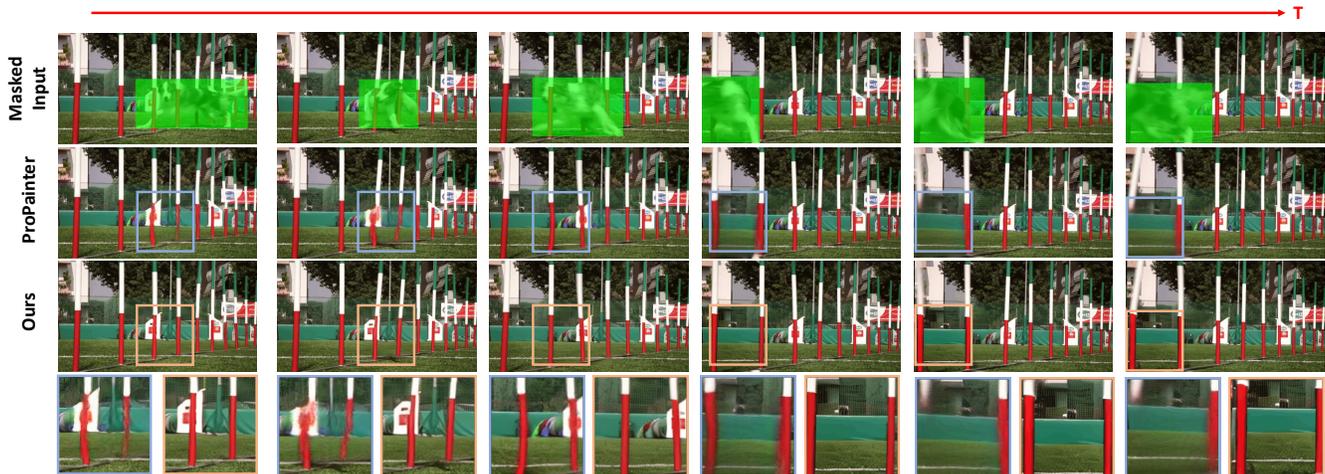
Figure 1: Inpainting results of the proposed FFF-VDI and the flow propagation-based ProPainter. When the target object is frequently occluded or structurally difficult to track, large and rough bounding box masks are advantageous for editing.

the missing parts of the first frame's noise latent code. Finally, we fill the masked areas of the future frames with random noise and use the Deformable Noise Alignment (DNA) module to improve temporal consistency and minimize distortion at the noise latent level. By completely filling the first frame latent information with information from future frames and using these as conditional features, we can fine-tune the pre-trained image-to-video diffusion model to generate much more natural inpainting videos.

This approach has the advantage of significantly reducing dependency on the accuracy of completed optical flow compared to traditional methods (Gao et al. 2020; Zhang, Fu, and Liu 2022b,a; Zhou et al. 2023) that propagate all frame pixels or feature information using optical flow. Figure 1 shows the object removal performance of ProPainter (Zhou et al. 2023), which is based on optical flow propagation, and the proposed FFF-VDI. As shown in the figure, the existing method applies optical flow propagation to all frames, resulting in errors accumulating and causing texture blurring or distortion in the resulting video. However, the proposed method applies noise latent propagation only to the first frame and utilizes the strong temporal consistency performance of the video diffusion model for the subsequent frames, resulting in much more consistent and natural videos. Furthermore, our method considers the actual pixel information of the areas occluded by the mask. Traditional video diffusion models prioritize maintaining temporal consistency between neighboring frames. However, the proposed method brings in future latent information to reconstruct the actual pixel information in the masked areas of the first frame. As a result, temporal consistency is maintained even when the mask moves or previously occluded areas are revealed. Furthermore, to address the hallucination effects inherent in traditional diffusion models without using text guidance, the proposed FFF-VDI inference process applies DDIM inversion (Mokady et al. 2023). We conduct DDIM inversion (Mokady et al. 2023) to the entire video frame, propagating the inverted noise to the first frame. This inverted noise generates only the actual erased areas during the denoising process, thus minimizing object hallucination compared to traditional methods that use random noise.

We conduct various comparative experiments to demonstrate that the proposed FFF-VDI outperforms previous methods in both video completion and object removal across different scenarios. Furthermore, we show that our method has robust performance with large and rough masks for object removal, compared to existing optical flow propagation-based methods.

Related Work

Flow propagation-based approaches. Flow propagation-based approaches leverage the relatively simpler task of flow completion to assist in the more complex task of RGB content filling. Flow-based video inpainting methods generally follow a three-phase process: flow completion, content propagation, and content generation. Various methods have been proposed to improve each phase. For example, FGVC (Gao et al. 2020) integrates gradient propagation during the content propagation phase, and E2FGVI (Li et al. 2022) introduces an end-to-end flow completion module. Additionally, FGT (Zhang, Fu, and Liu 2022a) combines decoupled spatiotemporal attention with gradient propagation techniques from FGVC. ProPainter (Zhou et al. 2023) pushes the boundaries further by integrating dual-domain propagation. However, these methods often struggle with spatial misalignment due to flow propagation errors or suffer from diminished detail retention caused by repetitive resampling in attempts to achieve sub-pixel accuracy.

Diffusion-based approaches. Diffusion-based methods generate inpainted frames by leveraging the generative capabilities of the diffusion model and a module that maintains temporal consistency at the noise level. For example, AVID (Zhang et al. 2024) uses an image diffusion model to naturally fill each frame and applies a temporal consistency module during the inference stage to generate inpainted videos. CoCoCo (Zi et al. 2024) is similar to AVID but
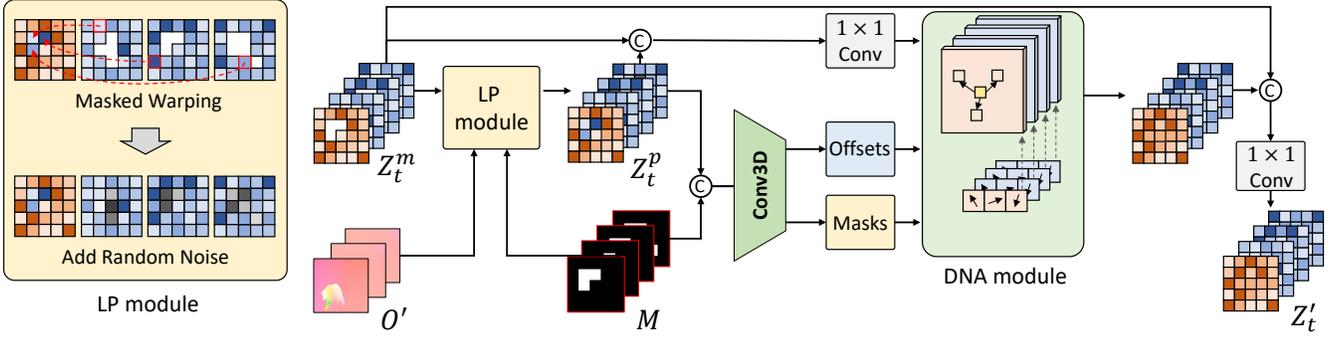
Figure 2: The overall training and testing pipeline structure of our FFF-VDI.

additionally introduces a motion capture module to maintain motion consistency in the generated video. However, these methods focus on improving perceptual temporal consistency and do not consider the actual pixel information in the masked areas.

Preliminaries

Video Diffusion Model. Most video-based diffusion models (Rombach et al. 2022; Blattmann et al. 2023; Ho et al. 2022) use a 3D-Unet to learn how to remove noise sequences randomly sampled from a Gaussian distribution. First, a pretrained 2D-VAE (Kingma and Welling 2013) encoder $E(\cdot)$ is used to extract the latent codes for each of the S frames of the input video, forming a video latent code sequence $\{Z_{0_i}\}_{i=1}^S$. Then, this latent sequence is concatenated along the temporal dimension to generate the video latent code z_0. In the forward diffusion procedure, Gaussian noise is gradually added to Z_0. Let T be the total number of time steps in the diffusion process and $\{\beta_t\}_{t=1}^T$ the noise scheduler. The noisy latent code Z_t at time step t is expressed as follows:

$$Z_t = \sqrt{\bar{\alpha}_t}Z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I), \quad (1)$$

where $\bar{\alpha}_t = \prod_{i=1}^t \alpha_t, \alpha_t = 1 - \beta_t$ and ϵ is the adding noise code sampled from standard normal distribution $\mathcal{N}(0, I)$. As a result, the model aims to predict the noise ϵ from Z_t and t. If the parameters of the 3D-UNet are denoted as θ, the final objective function $\mathcal{L}$ is expressed as follows:

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right]. \quad (2)$$

In the inference stage, if the video latent code is denoted as Z_0, noise is iteratively removed at each time step using the trained 3D-UNet. Finally, the video frames are reconstructed by the 2D-VAE decoder $D(\cdot)$.

Proposed Approach

FFF-VDI Trainig Stage. Figure 2 (a) shows the overall architecture of the proposed FFF-VDI. The FFF-VDI aims to generate new completed video frames $I' = \{I'_1, I'_2, \ldots, I'_S\}$ using the masked S input frames $I^m = \{I^m_1, I^m_2, \ldots, I^m_S\}$.

In the FFF-VDI training stage, we pass each frame of I through the 2D-VAE encoder E to generate the video latent code $L_0 \in \mathbb{R}^{S \times C \times H \times W}$ and add time step noise to obtain the noisy latent code L_t. Then, we pass only the unmasked parts of the randomly masked frames I^m through E to generate the masked conditional latent code L^m. In other words, if the mask frames downsampled to the latent size are denoted as $M = \{M_1, M_2, \ldots, M_S\}$, the conditional latent code of the i-th frame L^m_i is expressed as $L^m_i = E(I^m_i) \odot (1 - M_i)$. Next, following the structure of Video LDM (Rombach et al. 2022), we proceed with the process of merging the conditional latent and the noisy latent. As shown in Figure 2, the masked conditional latent L^m is concatenated with L_t along the channel dimension and then merged into the masked noisy latent code Z^m_t through a 1×1 convolution layer.

Unlike existing methods that use pixel-level flow propagation for all frames, FFF-VDI applies noise latent level optical flow propagation only in the direction toward the first frame. To achieve this, we predict the masked optical flow O^m from I^m as shown in Figure 2 and apply a flow completion module to convert it into the completed optical flow O'. The flow completion module is widely used in traditional pixel propagation methods (Zhang, Fu, and Liu 2022a; Kang, Oh, and Kim 2022; Li et al. 2022; Zhang, Fu, and Liu 2022b; Zhou et al. 2023), and we ultimately use the pretrained flow completion module proposed by ProPainter (Zhou et al. 2023). O' and Z^m_t are used as inputs to the First Frame Filling (FFF) module to fill the first frame masked noise latent $Z^m_{t_1}$. More specifically, the noise latent codes $Z^m_{t_{2:S}}$ from all frames except the first are propagated to $Z^m_{t_1}$ based on the optical flow. As a result, the FFF module generates the completed first frame noise latent Z'_t.

Finally, Z'_t is used as the input to the pre-trained 3D-UNet for the denoising process. The objective function of FFF-VDI is the same as in Eq (2). We use the pre-trained image-to-video 3D-UNet from the Stable Video Diffusion (Blattmann et al. 2023). Therefore, to retrain the model for the video inpainting task, we fine-tune some layers of the 3D-UNet, as shown in Figure 2

FFF-VDI Testing Stage. In Figure 2 (b), we illustrate the

Figure 3: The structure of the proposed FFF module. First, the FFF module propagates the latent noise information from each frame to the first frame's latent noise to fill the masked areas. Next, deformable convolution is applied to reconstruct the latent-level distortions and structural information.

testing stage of the proposed FFF-VDI, which is the diffusion inference process. During the generating stage, most video diffusion models use randomly initialized Gaussian noise, drawn from a normal distribution N, as the initial conditional latent. However, in the denoising process, when there is no text guidance, unwanted objects may appear in the inpainting task due to hallucination effects. To address this issue, instead of providing random noises for the entire areas during the denoising process, we use controlled noises to prevent the generation of unwanted objects by fully utilizing the given pixels. Specifically, we generate the original noisy latent from the masked video using DDIM inversion and propagate it to the first frame latent through the FFF module.

To explain in more detail, we first pass the masked video through a 2D-VAE encoder to generate the masked conditional latent code L^m. Next, using the DDIM inversion process, we transform L^m into the inverted noise latent code L_T^i at time step T. Then, we merge different noise $N \sim \mathcal{N}(0, I)$ and L_T^i with L^m through the same convolution layer to generate Z_T and Z_T^i, respectively. Since FFF-VDI completes the first frame latent code using only the latent information from other frames, we use the newly concatenated latent code, formed by concatenating Z_{T_1} and $Z_{T_{2:S}}^i$ along the temporal axis, as the input for FFF module. Finally, we convert the latent code that has passed through the FFF module into the final denoised latent code using the standard diffusion denoising process. Similar to existing video inpainting evaluation processes, we composite the generated masked areas with the masked original video to produce the final video.

First Frame Filling Module. Figure 3 shows the structure of the proposed FFF module, which includes the Latent Propagation (LP) module and the Deformable Noise Alignment (DNA) module, First, LP module use the downsampled optical flow O' and the downsampled mask map M to warp the noisy latent information $Z_{t_{2:S}}^m$ from other frames to the first frame $Z_{t_1}^m$. To preserve the original first frame information, the latent information is propagated only to the masked regions of $Z_{t_1}^m$, which correspond to the unmasked regions of $Z_{t_{2:S}}^m$. Therefore, the warping process can be expressed as follows:

$$Z_{t_1}^p = (1 - M_1) Z_{t_1}^m + M_1 \sum_{i=2}^{S} \mathcal{W}\left((1 - M_i) Z_{t_i}^w, O_{i \to 1}'\right),$$

(3)

where $\mathcal{W}(\cdot)$ denotes warping operation. Additionally, we sequentially apply optical flow to warp the latent from the i-th frame to the first frame. In other words, the flow $O_{i \to 1'}'$, which warps from the i-th frame to the first frame, is composed of the sequential set $\{O_{i \to i-1}', O_{i-1 \to i-2}', \ldots, O_{2 \to 1}'\}$. Next, we add random noise to the masked regions of $Z_{t_{2:S}}^m$ and combine it with the first frame noisy latent to generate the propagated Z_t^p. This process can be expressed as follows:

$$Z_t^p = Z_1^p \| \left(Z_{t_{2:S}}^m + \epsilon M_{2:S}\right), \epsilon \sim \mathcal{N}(0, I),$$

(4)

where $\|$ is the concatenation operation along the temporal axis.

Unlike the first frame where flow propagation is applied, we fill the masked areas of future frames with random noise, which can lead to a decrease in temporal consistency during the diffusion denoising process. Therefore, we add a noise refinement process to maintain the temporal consistency of the generated noisy latent. For this, we introduce a DNA process using a deformable convolution network (DCN). Although a similar approach is proposed in E2FGVI (Li et al. 2022), our model differs in that it does not use optical flow for DCN offset prediction. This minimizes the distortion and consistency errors in temporal noisy latents caused by incorrect flow completion. Instead, as shown in Figure 3, the DNA module applies 3D convolution layers to directly learn DCN offsets from the noisy latent. As a result, we combine the refined noisy latent with Z_t^m to generate the final noisy latent Z_t'. Specifically, we first concatenate the previously generated Z_t^p and the mask M and pass them through a stack of 3D convolution layers. We then separate this feature to generate the offset o and the DCN modulation mask m. m adjusts the contribution of each noisy latent pixel in the DCN. Our DNA process is expressed as follows:

$$\hat{Z}_{t_s} = \mathcal{R}\left(\mathcal{D}\left(\hat{Z}_{t_{s+1}}, o_{s \to s+1}, m_{s \to s+1}\right), \hat{Z}_{t_s}\right),$$

(5)

Methods	Publication	YouTube-VOS				DAVIS			
		PSNR↑	SSIM↑	VFID↓	E_{warp}^*↓	PSNR↑	SSIM↑	VFID↓	E_{warp}^*↓
FGVC (Gao et al. 2020)	ECCV 2020	29.67	0.9403	0.064	1.163	30.80	0.9497	0.165	1.571
STTN (Zeng, Fu, and Chao 2020)	ECCV 2020	32.34	0.9655	0.053	1.061	30.61	0.9560	0.149	1.438
TSAM (Zou et al. 2021)	CVPR 2021	30.22	0.9468	0.070	1.014	30.67	0.9548	0.146	1.235
FuseFormer (Liu et al. 2021)	ICCV 2021	33.32	0.9681	0.053	1.053	32.59	0.9701	0.137	1.349
ISVI (Zhang, Fu, and Liu 2022b)	CVPR 2022	30.34	0.9458	0.077	1.008	32.17	0.9588	0.189	1.291
FGT (Zhang, Fu, and Liu 2022a)	ECCV 2022	32.17	0.9599	0.054	1.025	32.86	0.9650	0.129	1.323
E^2FGVI (Li et al. 2022)	CVPR 2022	33.71	0.9700	0.046	1.013	33.01	0.9721	0.116	1.289
ProPainter (Zhou et al. 2023)	ICCV 2023	34.43	0.9735	0.042	0.974	34.47	0.9776	0.098	1.187
FFF-VDI (Ours)		**35.06**	**0.9812**	**0.031**	**0.937**	**35.03**	**0.9834**	**0.075**	**1.009**

Table 1: Quantitative comparison on the YouTube-VOS and DAVIS datasets. E_{warp}^* denotes $E_{warp}\left(\times 10^{-3}\right)$.

where $\mathcal{D}(\cdot)$ denotes deformable convolution, and $\mathcal{R}(\cdot)$ denotes the convolution layers that fuse the aligned and current features. Additionally, $\hat{Z}_t$ is the feature where Z_t^m and Z_t^p are merged. As a result, we combine the refined noisy latent with Z_t^m to generate the final noisy latent Z_t'.

Experiments

Datasets. To fairly compare previous state-of-the-art models with the proposed FFF-VDI, we use the YouTube-VOS (Xu et al. 2018) training set as the training data. Additionally, for model evaluation, we use the widely known YouTube-VOS (Xu et al. 2018) and DAVIS (Perazzi et al. 2016) test sets as evaluation datasets. The YouTube-VOS test set consists of 508 video clips, and the DAVIS test set consists of 90 video clips. For the DAVIS test set, we follow the approach of ProPainter (Zhou et al. 2023) and E2FGVI (Li et al. 2022), using 50 video clips for evaluation.

Training Details and Evaluation Metrics. In this paper, we use the optical flow prediction model RAFT (Teed and Deng 2020) to extract optical flow. Additionally, we use the pre-trained flow completion model proposed by ProPainter (Zhou et al. 2023) to inpaint the masked optical flow. To ensure a fair comparison with prior studies, all videos are resized to 432×240. However, due to the specific input size required by the 3D-UNet of the Video LDM (Rombach et al. 2022), we pad the input frames to change its size to 448×256. Also, we follow previous methods (Li et al. 2022; Liu et al. 2021; Zhou et al. 2023) to randomly generate stationary and object masks to simulate masks for video completion and object removal tasks. Our pre-trained image-to-video diffusion model follows the official implementation of Stable Video Diffusion (Blattmann et al. 2023), and we fine-tune the model by freezing all layers of the 3D-UNet except for the temporal transformer block. In this paper, we set the batch size to 4, the initial learning rate to 10^{-5}, and train for a total of 100,000 iterations with Adam (Kingma and Ba 2014) optimizer. Our method is implemented using the PyTorch framework and trained on four NVIDIA RTX A6000 GPUs.

For model evaluation, we use PSNR and SSIM (Wang et al. 2004), which are traditionally used in video inpainting tasks. Additionally, to assess the temporal consistency and smoothness of the resulting video sequences, we use the flow warping error E_{warp} (Lai et al. 2018) as an evaluation metric. Furthermore, following recent video inpainting researches (Liu et al. 2021; Li et al. 2022; Zhou et al. 2023), we report the VFID (Wang et al. 2018) score to measure the perceptual similarity between the output and ground truth videos.

Results

Quantitative results. In Table 1, we quantitatively compare our proposed FFF-VDI with state-of-the-art methods on the YouTube-VOS and DAVIS datasets. The removal mask maps for testing are the same as the official test mask maps provided by ProPainter (Zhou et al. 2023). As shown in the table, our method significantly outperforms all existing methods in the video completion task. Notably, we demonstrate greater performance improvement on the perceptual metric VFID than on traditional metrics like PSNR and SSIM. This is because traditional metrics tend to give higher scores to blurry textures rather than natural textures. Therefore, traditional metrics fail to properly evaluate the generation quality of blurry images resulting from incorrect optical flow predictions in previous flow propagation-based methods. Furthermore, flow propagation-based methods typically demonstrate high time consistency and excellent E_{warp} performance because they use flow to predict the next frame. However, due to their limited generative capability, they cannot effectively reconstruct missing pixel information. Therefore, the results in Table 1 show that the proposed FFF-VDI can effectively integrate the superior generative ability and time consistency of the image-to-video diffusion model into the video inpainting task without direct pixel propagation.

Qualitative results. Figure 4 shows the qualitative evaluation results of the previous state-of-the-art method, ProPainter (Zhou et al. 2023), and the proposed FFF-VDI across various scenarios. For the sake of a fair comparison, both ProPainter and FFF-VDI use the same number of reference frames. As mentioned earlier, visual quality is often not adequately reflected in quantitative metrics, making it crucial to compare qualitative evaluations. Therefore, in Figure 4, we present (a) random mask video inpainting results and (b) rough rectangular mask object removal results. As illustrated, the proposed model demonstrates significantly higher performance in propagating actual contextual information to the occluded areas and generating new content that cannot be observed, compared to existing methods. Particularly, the proposed method achieves natural video synthesis with high
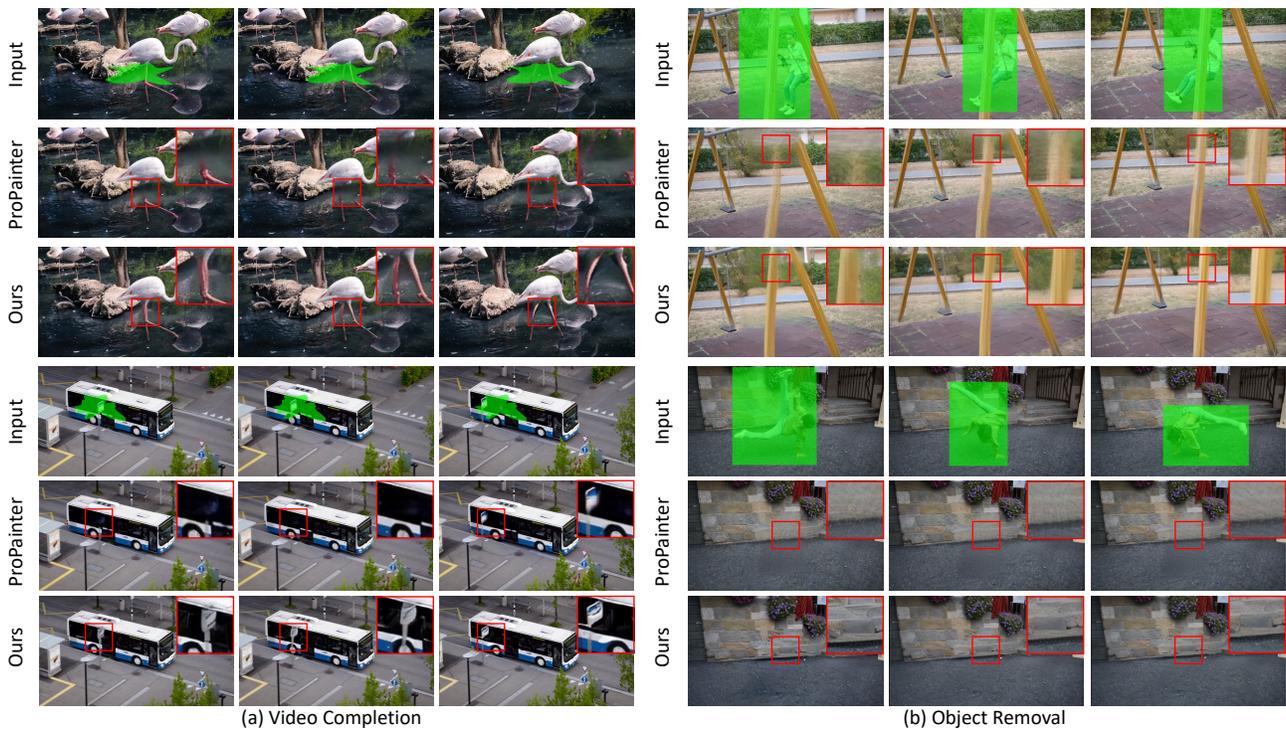
(a) Video Completion

(b) Object Removal

Figure 4: Qualitative comparisons on both video completion and object removal. The proposed FFF-VDI demonstrates robust video inpainting performance in masked areas compared to the existing flow propagation-based model, ProPainter.
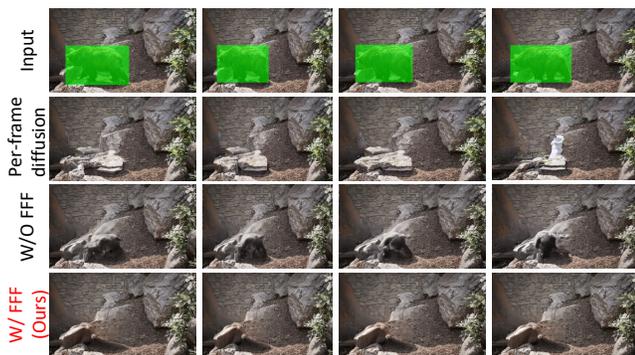

Figure 5: Visualization results with and without the proposed FFF module. Per-frame diffusion is the result of inpainting each frame with stable diffusion.

Figure 6: Visualization results with and without the DNA module.

time consistency and fewer distortions in areas where flow propagation is not possible, due to its lower dependence on optical flow.

Ablation Analysis

Effect of first frame filling. Table 2 and Figure 5 demonstrate the effectiveness of the proposed FFF module. First, the second row of Figure 5 shows the results of applying Stable Diffusion (Rombach et al. 2022) frame-by-frame to each masked area, where we perform inpainting with an empty text input. While the diffusion-based model provides natu-

ral fill-in images on a frame-by-frame basis due to its powerful generative capabilities, it exhibits very low temporal consistency quality. The third row shows our baseline model without the FFF module. In this case, even without applying latent propagation, the pre-trained video diffusion model's generative capabilities allow for the creation of a fairly temporally consistent video. However, the video generated with random noise does not reflect the actual context information of the initial frames. Therefore, as the mask moves and the real background behind the mask is revealed, perceptual in-
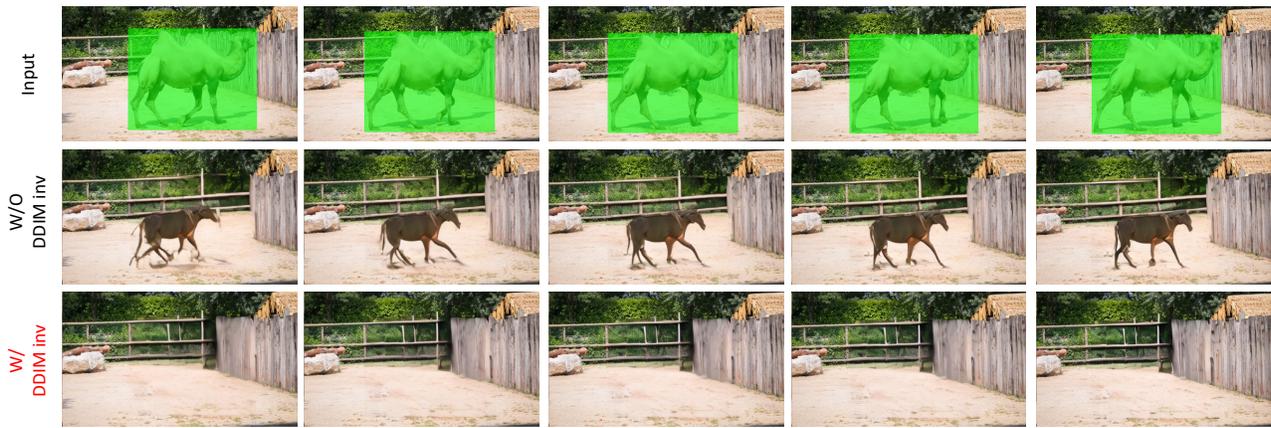
Figure 7: Comparison of generated images with and without DDIM inversion.

Method				YouTube-VOS				DAVIS			
Per-frame	FFF-VDI		DDIM inv	PSNR↑	SSIM↑	VFID↓	$E^*_{warp}\downarrow$	PSNR↑	SSIM↑	VFID↓	$E^*_{warp}\downarrow$
	LP	DNA									
✓				32.21	0.9532	0.063	1.592	30.59	0.9572	0.128	1.478
	✓			34.03	0.9783	0.043	1.142	34.11	0.9798	0.089	1.206
	✓	✓		35.01	0.9798	0.031	0.940	35.10	0.9814	0.074	1.034
	✓	✓	✓	35.06	0.9812	0.031	0.937	35.03	0.9834	0.075	1.009

Table 2: Comparison of random mask video inpainting performance based on the proposed contribution combinations.

consistencies arise between past and future frames. In other words, it is challenging for the existing diffusion model to model long-range temporal consistency, which is a major reason why applying video diffusion models to video inpainting tasks is difficult. However, the proposed FFF-VDI pre-completes the accurate background information of the first frame using the latent codes from the other frames, thus generating the most natural background video, as shown in the last row of Figure 5. These results are also well-reflected in the quantitative analysis results in Table 2.

Effect of deformable noise alignment. Table 2 and Figure 6 demonstrate the effect of the DNA module. As shown in Figure 6, when the mask size is very large, the reference frame provides limited visual cues, resulting in images with unnatural structures. Although ProPainter uses optical flow-based deformable convolution, it fails to reconstruct structural information when the flow information is also very limited, as seen in Figure 6. The proposed deformable alignment minimizes distortion in the denoised result images at the noise latent level instead of relying on optical flow, aiding in the generation of natural images.

Effect of DDIM inversion. Figure 7 compares the generated image results of FFF-VDI with and without the DDIM inversion process. As shown in the second row, due to the powerful generative capabilities of the pre-trained video diffusion model, hallucination effects occur, generating unwanted objects when the target mask size is large. However, as shown in the third row, when noise latent information is provided through DDIM inversion during the inference stage, FFF-VDI is able to generate normal videos. This trend is also reflected in the quantitative analysis results in Table 2, al-

though the effect of DDIM inversion is less significant due to the relatively smaller mask size used by ProPainter. In fact, most diffusion models use text guidance to address hallucination effects, but providing text guidance for backgrounds is challenging in video inpainting tasks. Therefore, the proposed FFF-VDI effectively resolves the issues of existing video diffusion models for video inpainting tasks and integrates the model effectively.

Conclusion

We introduce FFF-VDI, a novel video diffusion inpainting model based on First Frame Filling. Unlike traditional propagation-based models, FFF-VDI uses noise latent propagation for the first frame and leverages the temporal consistency of fine-tuned image-to-video diffusion models for subsequent frames. FFF-VDI addresses two major challenges in video inpainting. It ensures natural video generation when masks move by reconstructing the actual appearance behind the mask and minimizes object hallucination by using DDIM inversion. Comparative experiments show that FFF-VDI outperforms previous methods in video completion and object removal, especially with large and rough masks. This makes FFF-VDI a robust and effective solution for complex video inpainting tasks.

Acknowledgements

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Gao, C.; Saraf, A.; Huang, J.-B.; and Kopf, J. 2020. Flow-edge guided video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 713–729. Springer.

Gu, B.; Yu, Y.; Fan, H.; and Zhang, L. 2023. Flow-guided diffusion for video inpainting. *arXiv preprint arXiv:2311.15368*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Kang, J.; Oh, S. W.; and Kim, S. J. 2022. Error compensation framework for flow-guided video inpainting. In *European conference on computer vision*, 375–390. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M.-H. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 170–185.

Li, Z.; Lu, C.-Z.; Qin, J.; Guo, C.-L.; and Cheng, M.-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*, 17562–17571.

Liu, R.; Deng, H.; Huang, Y.; Shi, X.; Lu, L.; Sun, W.; Wang, X.; Dai, J.; and Li, H. 2021. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14040–14049.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 724–732.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Liu, G.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 585–601.

Zeng, Y.; Fu, J.; and Chao, H. 2020. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 528–543. Springer.

Zhang, K.; Fu, J.; and Liu, D. 2022a. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, 74–90. Springer.

Zhang, K.; Fu, J.; and Liu, D. 2022b. Inertia-guided flow completion and style fusion for video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5982–5991.

Zhang, Z.; Wu, B.; Wang, X.; Luo, Y.; Zhang, L.; Zhao, Y.; Vajda, P.; Metaxas, D.; and Yu, L. 2024. AVID: Any-Length Video Inpainting with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7162–7172.

Zhou, S.; Li, C.; Chan, K. C.; and Loy, C. C. 2023. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10477–10486.

Zi, B.; Zhao, S.; Qi, X.; Wang, J.; Shi, Y.; Chen, Q.; Liang, B.; Wong, K.-F.; and Zhang, L. 2024. CoCoCo: Improving Text-Guided Video Inpainting for Better Consistency, Controllability and Compatibility. *arXiv preprint arXiv:2403.12035*.

Zou, X.; Yang, L.; Liu, D.; and Lee, Y. J. 2021. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16448–16457.