Advanced Video Inpainting Using Optical Flow-Guided Efficient Diffusion

Bohai Gu[1,2*]   Hao Luo[2,†]   Song Guo[1,†]   Peiran Dong[1]

[1] Hong Kong University of Science and Technology [2] DAMO Academy, Alibaba Group

Abstract

*Recently, diffusion-based methods have achieved great improvements in the video inpainting task. However, these methods still face many challenges, such as maintaining temporal consistency and the time-consuming issue. This paper proposes an advanced video inpainting framework using optical **Flo**w-guided **E**fficient **D**iffusion, called FloED. Specifically, FloED employs a dual-branch architecture, where a flow branch first restores corrupted flow and a multi-scale flow adapter provides motion guidance to the main inpainting branch. Additionally, a training-free latent interpolation method is proposed to accelerate the multi-step denoising process using flow warping. Further introducing a flow attention cache mechanism, FLoED efficiently reduces the computational cost brought by incorporating optical flow. Comprehensive experiments in both background restoration and object removal tasks demonstrate that FloED outperforms state-of-the-art methods from the perspective of both performance and efficiency.*

1. Introduction

Video inpainting aims to predict corrupted regions by filling them with contextually appropriate and temporally coherent content, which plays an essential role in computer vision, particularly in background restoration (BR) [17, 23, 34] and object removal (OR) [4, 19, 33].

Although conventional transformer-based approaches [5, 15, 37] have made great progress on standard benchmarks [20, 31], they encounter difficulties when requiring the synthesis of new content not present in existing frames. Currently, diffusion models [36, 38] have demonstrated significant prowess in text-guided video inpainting. However, temporal consistency remains an area with substantial scope for further improvement. And current methods struggle to generate satisfying results on BR and OR tasks, which is
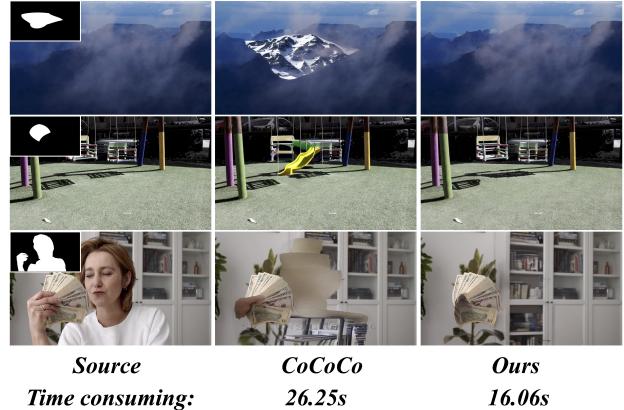

Figure 1. Superiority of FloED on background restoration and object removal.

probably caused by the training priors of powerful base models. As demonstrated in Fig. 1, contents inpainted by CoCoCo [38] lacks harmony with its surroundings. This disharmony can be alleviated by providing motion guidance that is aligned with the surroundings. Optical flow, a key modality for motion information, provides significant guidance and enhances temporal consistency. Notably, integrating optical flow into video inpainting requires additional operations, including flow completion and effective incorporation, which incur extra computational costs. And diffusion inherently suffers from efficiency due to the multi-step denoising process. Consequently, when leveraging optical flow, it is essential to consider efficiency enhancements tailored to the multi-step characteristics of diffusion models.

Based on the above analysis, we propose an advanced video inpainting framework using optical **Flo**w-guided **E**fficient **D**iffusion, called FloED. By leveraging motion information, our approach aims to enhance the reliability and performance of diffusion-based techniques in BR and OR applications. Initially, built on the large-scale T2I model, we inflate the origin layers and integrate pre-trained motion module [7] to establish a basic text-to-video framework. Then we fine-tune the motion module to effectively adapt it for the video inpainting domain. By adopting this

framework as the primary inpainting branch, our innovations focus on three key aspects: (1) We design a tailored flow branch that completes corrupted flow while maintaining consistent channel numbers with the primary branch. Next, we integrate a multi-scale flow adapter that incorporates flow features into the decoder blocks of the primary U-Net architecture, enabling FloED to utilize motion information more effectively. (2) Building on the observation that adjacent latents share similar motion patterns [35] and recognizing that the essence of diffusion models involves fundamentally multi-step sampling processes, we introduce a training-free latent interpolation technique. This method leverages warping operation guided by optical flow to effectively accelerate the multi-step denoising process during early denoising stage. Furthermore, by incorporating a flow attention cache mechanism during the remain denoising stage as a complementary speed-up solution, we efficiently minimize the additional computational burden typically introduced by flow adapters and flow branches. (3) Recognizing that state-of-the-art image inpainting models significantly outperform video inpainting diffusion models, we utilize an anchor frame strategy to enhance the quality of video inpainting outcomes.

Currently, there is no comprehensive benchmark for evaluating diffusion-based generative approaches in video inpainting. This deficiency presents a substantial challenge, as it limits the ability to rigorously assess and compare the efficacy of various inpainting methodologies. To bridge this gap, we have developed an extensive benchmark that meticulously encompasses both BR and OR tasks. Our main contributions are as follows:

- We propose a dedicated dual-branch architecture that incorporates motion guidance with a multi-scale flow adapter, thereby enhancing temporal consistency and overall quality.
- We introduce a training-free latent interpolation technique that leverages optical flow to speed up the multi-step denoising process. Complemented by a flow attention cache mechanism, FloED efficiently reduces the additional computational costs introduced by the flow.
- We conducted extensive experiments, including both quantitative and qualitative evaluations, to validate that FloED demonstrates superiority over other state-of-the-art methods in both BR and OR tasks.

2. Related Work

2.1. Conventional Video Inpainting

Conventional video inpainting mainly focuses on two types of tasks: object removal and background restoration. Object removal [4, 19, 33] aims to eliminate unwanted objects from video frames. And background restoration [17, 23, 34] involves seamlessly reconstructing missing region of the background with coherent content.

Given the fact that it's more simple to complete flows instead of directly filling masked regions [32], conventional transformer-based models [15, 35, 37] primarily rely on optical flow to propagate features or pixels for better inpainting outcomes. Specifically, E²FGVI [15] introduces an end-to-end framework with flow-guided feature propagation. FGT [35] combines decoupled spatiotemporal attention with a flow-guided content propagation. ProPainter [37] advances this field by merging dual-domain propagation with a mask-guided transformer.

Although propainter [37] has shown significant improvements on standard benchmarks [20, 31], it still struggles with generating coherent new content that isn't present in existing frames. There is a pressing need to develop more robust solutions.

2.2. Diffusion-based Video Inpainting

In recent years, diffusion models [9, 22] have revolutionized the field of content generation, showcasing exceptional abilities in creating highly realistic outputs [2, 21]. The development of text-to-video (T2V) [7, 10] and image-to-video (I2V) [3, 30] diffusion generative models, has paved the way for innovations in text-guided video inpainting. Models like AVID [36] and CoCoCo [38], have highlighted the routine practice of employing pre-trained Stable Diffusion Inpainting models [2] equipped with motion modules [7] for text-guided video inpainting. Specifically, AVID [36] introduce zero-shot generation pipeline with a structure guidance module. CoCoCo [38] implements enhanced attention mechanisms to address text-video alignment and ensure motion consistency across frames. While these advancements have contributed to progress, existing diffusion-based video inpainting methods still exhibit significant room for improving temporal consistency and they perform suboptimally on both OR and BR tasks. Compared with them, FloED showcases distinct advantages over temporal consistency and overall quality.

3. Preliminaries

LDM [21] leverages a pre-trained Variational Autoencoder (VAE) to operate in the latent space instead of pixel space. The diffusion forward process is imposing noise on a clean latent $\mathbf{z}_0$ for T times. A property of the forward process is that it admits sampling $\mathbf{z}^t$ at random timestep t:

$$q(\mathbf{z}^t|\mathbf{z}^0) = \mathcal{Q}(\mathbf{z}^0, t) = \mathcal{N}(\mathbf{z}^t; \sqrt{\bar{\alpha}_t}\mathbf{z}^0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (1)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$, β_s is the variance schedule for the timestep s. The backward process applies a trained UNet ϵ_θ for denoising: $p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t))$, where distribution parameters μ_θ and Σ_θ are computed by the denoising model θ. To
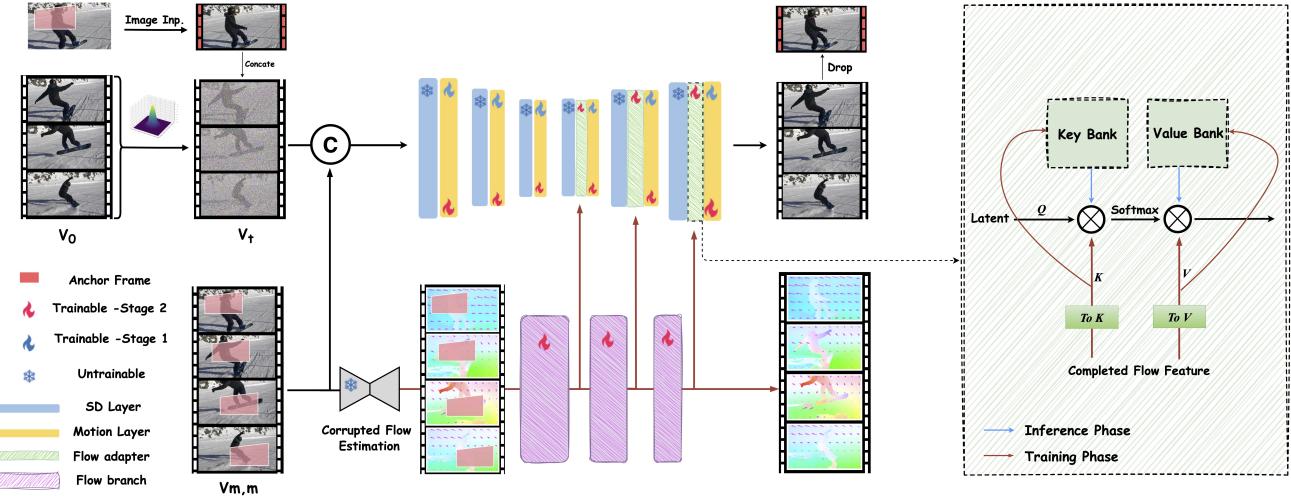
Figure 2. Overview of FloED. FloED employs a dual-branch architecture implemented through a two-stage training approach. In the first training stage, we focus exclusively on the upper branch, optimizing the motion layer to adapt specifically to the video inpainting domain. Subsequently, we introduce a dedicated flow branch complemented by a multi-scale flow adapter, which provides flow guidance covering upblocks of primary UNet. During the inference phase, we enhance efficiency by integrating the flow attention cache (right part).

train a conditional LDM, the objective is given by:

$$\mathcal{L}_{\text{diff}} = \arg\min_{\theta} \mathbb{E}_{\mathbf{z},\epsilon\sim\mathcal{N}(0,1),t,c} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c)\|_2^2\right], \quad (2)$$

where $\epsilon_\theta(\mathbf{z}_t, t, c)$ is the predicted noise based on $\mathbf{z}_t$, the time step t and the condition c. Once trained, we could leverage the deterministic sampling of DDIM [9] and PNDM [16] to denoise $\mathbf{z}_t$:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\hat{\mathbf{z}}_{t\to0}}_{\text{predicted `}\mathbf{z}_0\text{'}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(\mathbf{z}_t,t,c)}_{\text{direction pointing to } \mathbf{z}_t} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}, \quad (3)$$

where σ_t are hyper-parameters. The term $\mathbf{z}_{t\to0}^t$ represents the predicted $\mathbf{z}_0$ at time step t, For conciseness and to circumvent any potential confusion with the concept of optical flow, we subsequently refer to $\hat{\mathbf{z}}_{t\to0}$ as $\hat{\mathbf{z}}_0$. The precise formulation is as follows:

$$\hat{\mathbf{z}}_{t\to0} = \mathcal{P}(\mathbf{z}_t, \epsilon_\theta) = (\mathbf{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{z}_t,t,c))/\sqrt{\alpha_t}. \quad (4)$$

4. Methods

Building upon Stable Diffusion inpainting model [1], we integrate the motion module from Animatediff [7], adopting it as primary inpainting branch of our framework. As illustrated in Fig. 2, the training process of FloED can be divided into two stages. Initially, we adapt primary branch to video inpainting by fine-tuning the motion modules (Sec. 4.1).

Our approach then incorporates a dedicated flow branch to complete corrupted flows estimated from masked frames, alongside a multi-scale flow adapter that provides motion guidance for the primary inpainting branch (Sec. 4.2). To further enhance video inpainting results, we implement an anchor frame strategy to leverage the priority of the image inpainting diffusion model [2] (Sec. 4.3). Furthermore, we introduce a training-free denoising acceleration technique that leverages optical flow for latent interpolation, compensated with a flow attention caching mechanism in the reference phase. We substantially enhance efficiency while significantly reducing the additional computational overhead the flow introduces. (Sec. 4.4).

4.1. Domain Adaptation

Given an original video sequence $\mathbf{V_0} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{N-1}\} \in \mathbb{R}^{N\times3\times H\times W}$ and a binary mask sequence $\mathbf{m} = \{\mathbf{m}_0, \mathbf{m}_1, \ldots, \mathbf{m}_{N-1}\} \in \mathbb{R}^{N\times1\times H\times W}$, corrupted frames $\mathbf{V_m}$ are obtained by applying the Hadamard product as follows: $\mathbf{V_m} = \mathbf{V_0} \odot \mathbf{m}$. We aim to generate a set of spatiotemporally consistent inpainted outcomes with temporally coherent and contextually appropriate content in the corrupted area while ensuring out-of-mask areas are unchanged.

Following previous solutions [36, 38], FloED is built on Stable Diffusion (SD) inpainting model [1], equipped with motion module from AnimateDiff-v3 [7] for better temporal consistency. Notably, during the first stage, only the primary branch is taken into consideration, and we only optimize the motion layer for adapting to the video inpainting.
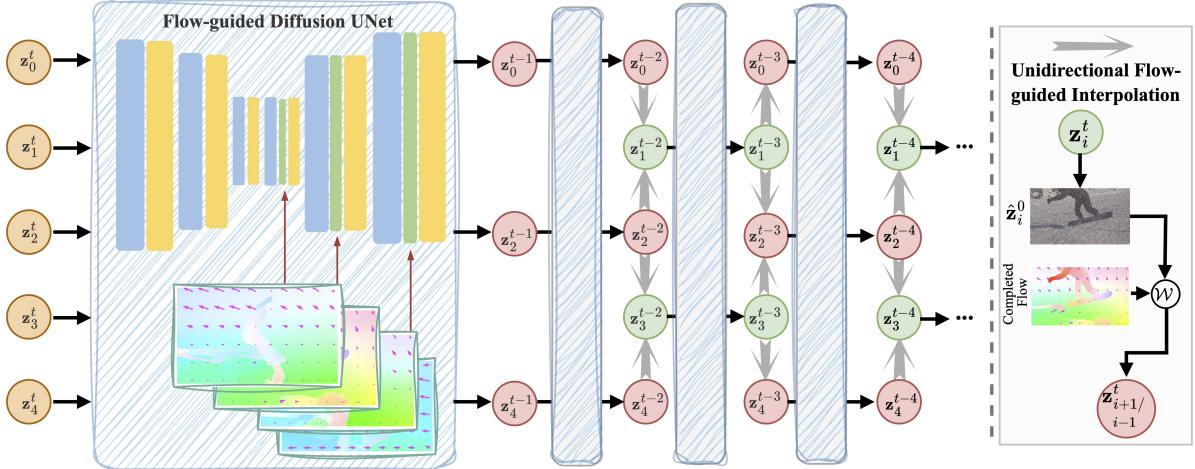
Figure 3. Illustration of flow-guided latent interpolation (left) and warping operation (right) during the denosing process.

4.2. Flow Guidance in Video Inpainting

Typically, for video inpainting methods [5, 15, 37] that utilize optical flow, it is typically necessary to first complete the corrupted flow because only an estimated flow from the masked frames is initially available. Specifically, we first employ a pre-trained flow estimator [24] to perform corrupted flow estimation from $\mathbf{V_m}$. Subsequently, we introduce a flow sub-branch composed of the first layer from each block of the primary branch, which is independent of the timestep. As illustrated in Fig. 2, the flow sub-branch completes the corrupted flow while maintaining consistent channel numbers with the latent representations of the corresponding frames in the primary branch. This approach enables more effective utilization of motion information by integrating multi-scale flow adapters into all upblocks of the primary UNet. As for the flow adapter, it consists of learnable modules with decoupled cross-attention mechanisms that embed flow features into the primary branch (Fig. 2), thereby providing robust motion guidance effectively. We choose to insert the flow adapter between the text attention and the motion layer because it can compensate for the information provided by the prompt and further supply temporal guidance to the motion layer. Consequently, in the second training phase, we also trained the motion layer simultaneously. An additional optical loss $\mathcal{L}_{\text{flow}}$ is introduced in the second stage: (F represents the ground truth flow.)

$$\mathcal{L}_{\text{sec}} = \mathcal{L}_{\text{diff}} + \lambda * \mathcal{L}_{\text{flow}}, \mathcal{L}_{flow} = \|\hat{F} - F\|_1 \quad (5)$$

4.3. Anchor Frame Strategy

Recognizing that state-of-the-art image inpainting models significantly outperform their video inpainting counterparts, we introduce an anchor frame strategy to enhance the quality of video inpainting results. As illustrated in Fig. 2, for

a given video sequence $\mathbf{V_0}$, we select an additional frame from the beginning of the sequence to serve as an anchor frame. We then utilize a pre-trained text-to-image (T2I) inpainting model [2] to reconstruct its corrupted region in advance. Subsequently, we concatenate the inpainted anchor frame with the noised video frames $\mathbf{V_t}$. This approach provides additional texture guidance to the video frames during the denoising process. After denoising, the anchor frame is discarded. This strategy leverages the superior performance of image inpainting models to improve the overall quality of video inpainting.

4.4. Efficient Inference for FloED

Based on multi-step sampling processes of diffusion, we further propose a training-free latent interpolation technique that leverages optical flow to speed up the denoising process. This approach is complemented by a flow attention cache mechanism during the inference phase.

Optical Flow Attention Cache Unlike the primary branch, the flow sub-branch is independent of timestep. During the inference phase, we utilize the flow sub-branch exclusively for flow completion in the first step and then use these completed flows for all subsequent steps. Regarding the flow adapter, it introduces additional computations by calculating the flow attention at every denoising step and multiple resolutions. To optimize this process, we establish a cache mechanism by computing the keys and values only during the first step and storing them in the memory bank (right part of Fig. 2). For the remaining steps, the cached keys and values are directly retrieved from the memory bank, eliminating the need for repeated calculations and enhancing efficiency.

Training-Free Denoising Speed-up Since adjacent feature latent exhibit similar motion patterns [35] and diffusion

"Fire burning in a fireplace, with a log burning on top of it." *"Forest with a stream running through it, surrounded by trees and plants."*
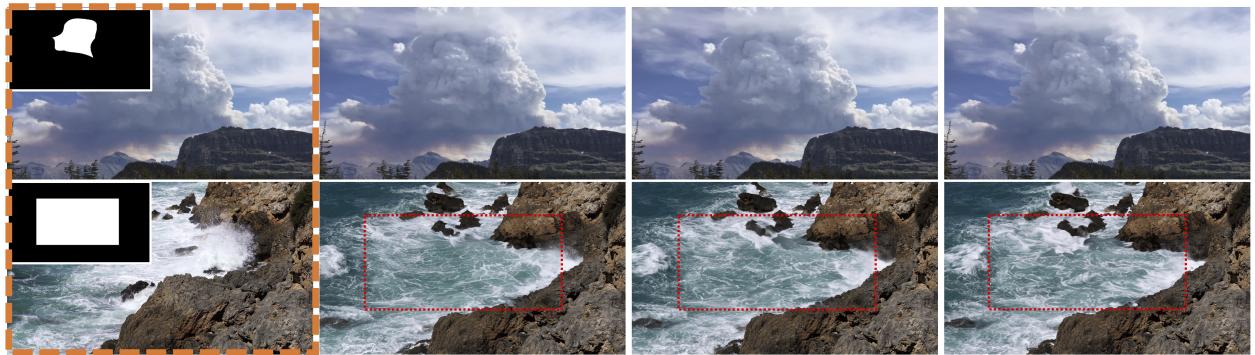

Figure 4. Qualitative Results. The orange dotted line indicates the input source.

models generate high-level content early in the denoising process [13], we aim to speed-up the denoising process by interpolating latent using the completed flow. Notably, this technique is entirely training-free.

Specifically, as illustrated in Fig. 3, the initial step involves performing the standard denoising process for completing flow and caching flow attention. Subsequently, starting from step $t - 1$, the noisy latent $\mathbf{z}$ is divided into two subsets based on parity. Latent interpolation process then follows a two-step alternating loop: even-indexed latents (shown in red) undergo denoising, while odd-indexed latents (shown in green) are obtained by warping operations using bi-directional optical flows. In the next step, only the interpolated latents (green) are denoised, and the red latents are generated through a similar warping process. Due to the negligible time cost of warping latent, the latency of denoising is halved by processing only half of the frame latent at each sampling timestep. Notably, the warping operation needs to be conducted at the $\mathbf{z}_0$ stage (as per Equation 4).

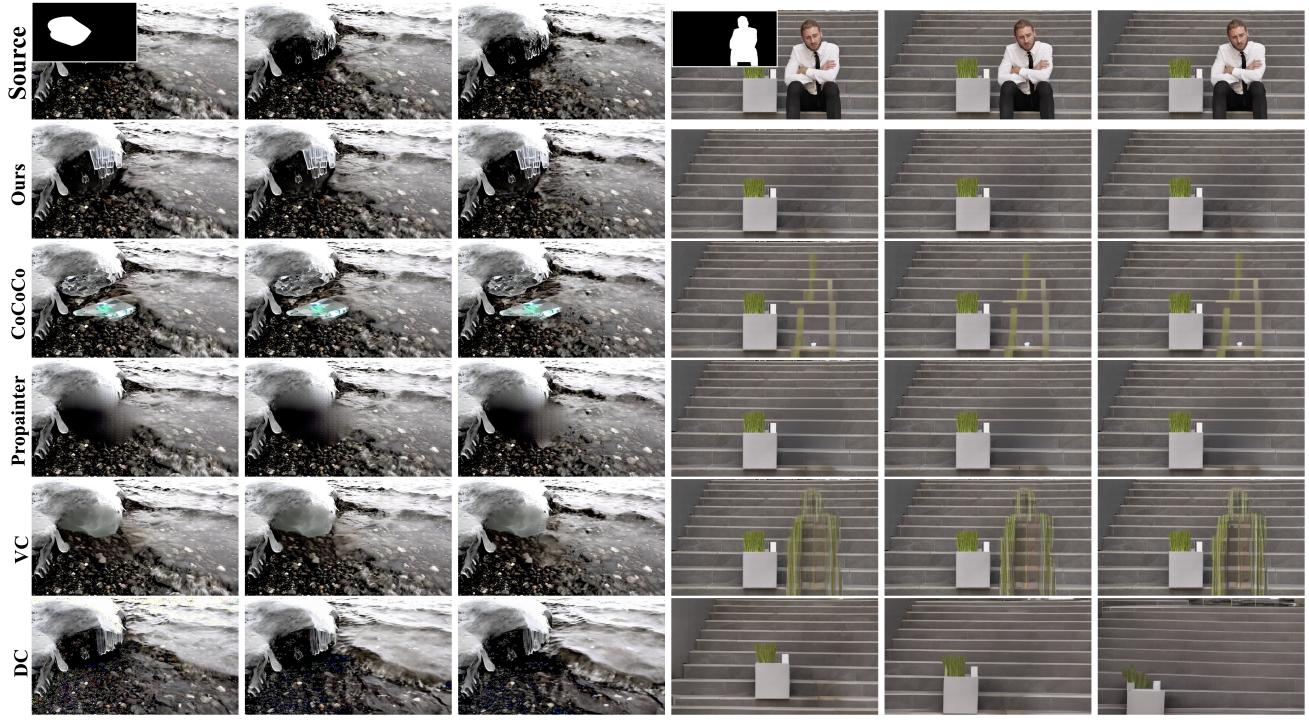Ideally, this process could be iterated until the final de-

noising step. However, since we are operating in latent space, completed optical flow provides only coarse-grained guidance, corresponding to the early stage of the diffusion process. Therefore, we restrict latent interpolation to the initial S denoising steps, during which the overall structure of the image is established [13]. Additionally, to minimize flow errors, we perform warp operations exclusively between adjacent frames. Furthermore, to mitigate potential occlusion issues in flow warping [11], we perform the copy-paste operation in each denoising step (see Algorithm in appendix).

5. Experiments

5.1. Implementation details.

FloED is built upon Stable Diffusion (SD) [1] and incorporates motion modules initialized from AnimateDiff v3 [7]. We employ RAFT [24] for flow estimation and SD v2 inpainting [2] for the anchor frame strategy.

For training, we utilize the video dataset from Open-

Figure 5. **Qualitative comparisons**. We compare our method against several approaches on BR and OR tasks.

Sora-Plan [14], which comprises over 400K high-quality videos sourced from Pexels and Pixabay platforms. During the training phase, we use 16 frames at a resolution of 256×256, and synthetic random masks are applied throughout the training process.

For the two-stage training process, the number of epochs is set to 5 and 30, respectively. And λ is set to 0.1 during the second stage. A batch size of 128 is achieved through gradient accumulation across eight NVIDIA A800 GPUs.

During inference, we empirically define the speed-up step S as 5 (25 steps). We have developed an evaluation benchmark comprising 100 previously unseen videos, with 50 designated for object removal (OR) and 50 for background restoration (BR). For BR task, we use synthetic random masks that focus on the background. For the OR task, object masks are obtained by applying Segment-Anything [12] (SAM) to each frame.

5.2. Qualitative Results

To comprehensively demonstrate the capabilities of our method, we evaluated FloED on real videos for object removal and background restoration tasks. As illustrated in Fig. 4, FloED effectively handles both OR and BR tasks across a wide range of mask sizes, including those covering more than half of the frame area. Our approach adeptly inpaints the specified regions with coherent and contextually appropriate content. Additionally, FloED exhibits excellent

temporal consistency throughout the video.

5.3. Comparisons

We present a comprehensive evaluation of our method compared to other open-source video inpainting solutions, including VideoComposer [26], CoCoCo [38], and Propainter [37]. Building upon the same image inpainting [2] results from anchor frames, we also assess the performance of the representative I2V model Dynamicrafter [29] as another comparison. We generate the corresponding video prompts for BR using VideoGPT [18], while manually creating appropriate background prompts for OR. Following the prior method [37], we applied a copy-paste approach to replicate the out-of-mask region across all methods except Dynamicrafter [29].

Qualitative comparisons. Fig. 5 (left) compares the performance of background restoration, while Fig. 5 (right) illustrates object removal. Compared to the representative transformer-based method Propainter [37], our approach generates more vivid and detailed results in object removal (right), whereas Propainter encounters difficulties when synthesizing new content (left). Regarding text-guided diffusion methods, CoCoCo [38] and VideoComposer [26] tend to produce hallucinated content that is inconsistent with the surrounding environment. When focusing on the shadow regions within the mask area (right), our method

demonstrates superior temporal consistency compared to Dynamicrafter [29].

Task	BR					OR
Metrics	PSNR ↑	SSIM ↑	VFID ↓	E_{warp}↓	TC ↑	TA ↑
DC	17.50	0.4157	0.934	6.17	0.985	22.44
VC	22.81	0.8614	0.193	3.43	0.987	21.30
Propainter	27.02	0.9057	0.129	**1.51**	**0.996**	20.94
CoCoCo	23.08	0.8694	0.165	3.73	0.991	21.97
Ours	**29.17**	**0.9441**	**0.118**	<u>2.83</u>	<u>0.994</u>	**22.49**

Table 1. **Quantitative comparisons** using different metrics, including PSNR, SSIM, FVID, Optical flow warping error, Temporal consistency (TC), and Text alignment (TA).

Quantitative comparisons. (a) Metric Evaluation. For background restoration, we employ PSNR [28], VFID [25], and SSIM [27] to quantify basic quality. Additionally, we assess temporal consistency using flow warping error [6, 13] in conjunction with Temporal Consistency (TC) [36]. TC is measured by the cosine similarity between consecutive frames in the CLIP-Image [8] feature space. For object removal, since ground truth data is unavailable for evaluating the aforementioned metrics, we utilize Text Alignment (TA) [8] as an evaluation metric, leveraging the CLIP score. For consistency, all metric evaluations were conducted at a resolution of 512×512. As shown in Tab. 1, our method outperforms other methods in PSNR, SSIM, VFID, and TA. Additionally, it surpasses other diffusion-based methods in flow warp error and TC, second only to Propainter by a small margin, due to the inherent generative characteristics of diffusion. Overall, FloED delivers excellent temporal consistency and high-quality outcomes compared to the state-of-the-art solutions. (b) User study. Since CLIP scores do not always align with human perception [36], we conducted a comprehensive user study. 12 annotators evaluated different inpainting results of 100 videos from 5 methods, assessing temporal consistency, content coherence, and overall quality. As illustrated in Fig. 6, our model was highly favored, achieving the highest scores in both BR
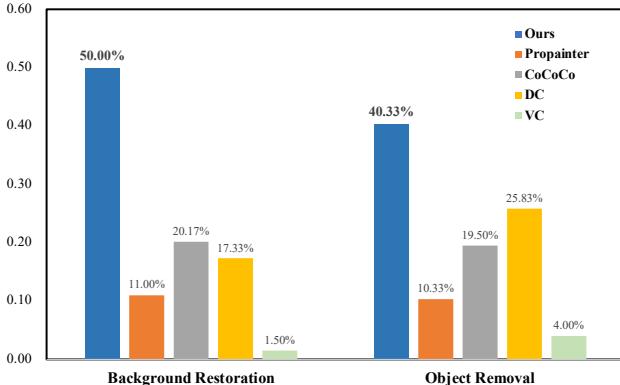

Figure 6. We conduct reliable **User Study** with randomized order to assess their performance across different inpaintng outcomes.

FA	AF	PSNR ↑	SSIM ↑	VFID ↓	E_{warp}↓	TC ↑
✗	✗	21.30	0.8435	0.246	4.15	0.989
✗	✔	25.34	0.9138	0.195	3.87	0.984
✔	✗	27.05	0.9255	0.170	3.03	0.989
✔	✔	28.71	0.9401	0.125	2.95	0.990
✔	✔	**29.17**	**0.9441**	**0.118**	**2.83**	**0.994**

Table 2. **Ablation study**. **FA**, **AF** stand for Flow Adapter and Anchor frame, respectively. All variants were trained using identical settings, except the last one with extended training (30 epochs).

(50.0%) and OR (40.33%), thereby demonstrating a consistent advantage over competing approaches.
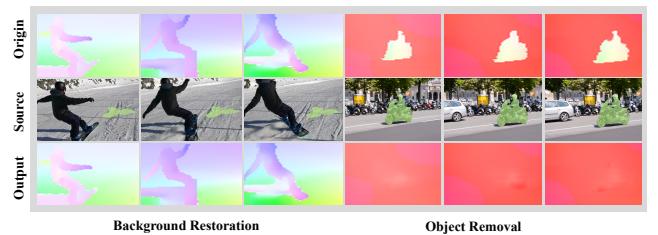
5.4. Ablation Study

Figure 7. **Optical Flow Completion**. Flows in first row are estimated by RAFT [24]. The completed flows provide motion guidance aligned with the mask surroundings.

Effect of flow guidance. We present the completed flow for the BR and OR tasks, as shown in Fig. 7. The corrupted flow is contextually inpainted with appropriate content that aligns with the surrounding environment while keeping the out-of-mask region intact, which serves as motion guidance for the primary inpainting branch. In Fig. 8, we further demonstrate the effects of flow guidance through two object removal examples. We compare two different variants of our model (3th and 5th rows of Tab. 2) to illustrate the impact of flow guidance on inpainting. The inpainted leaves with the flow adapter are more coherent with the surrounding environment (left), especially in the regions at the edges

Figure 8. Analysis of flow adapter. OR prompts are "Green leaves of the bush sway in the wind." and "A huge rock with flowing water in the background." respectively.

Figure 9. Analysis of anchor frame. BR prompt is "Grilled chicken wings on the barbecue, 8K."

Flow Branch	L.I.	Flow Cache	Time	SSIM↑
–	–	–	15.55	0.843
1st step	–	–	19.79	0.945
1st step	$2^{nd} \sim 6^{th}$	–	17.72 (↓ 10.5%)	0.944
1st step	$2^{nd} \sim 6^{th}$	$6^{th} \sim 25^{th}$	16.06 (↓ 18.8%)	0.944

Table 3. **Efficiency analysis.** L.I. represents the Latent Interpolation. The second row represents no flow-related computation involved. **As the closest counterpart, CoCoCo [38] takes 26.25s.**

of the object mask. Additionally, the repaired rocks exhibit greater temporal consistency (right) compared to outcomes without the flow adapter which demonstrates flow adapter ensures flow guidance effectively assists the model in generating content consistent with the environment, thereby enhancing temporal coherence and overall quality. Furthermore, as demonstrated in 3th and 5th rows of Tab. 2, the flow adapter significantly improves FloED's performance across all metrics, providing substantial benefits to video quality and temporal consistency.

Effect of anchor frame. As illustrated in Fig. 9, we investigate the impact of the anchor frame. Compared to the variant without the anchor frame strategy, incorporating an anchor frame introduces finer details, thereby enhancing the performance of FloED as an auxiliary strategy. Compared to the flow adapter, its effectiveness is somewhat inferior. As shown in 3th and 4th rows of Tab. 2, the variant incorporating the flow adapter achieves higher results than the one trained with anchor frame strategy.

Efficiency analysis. We conducted an efficiency analysis using a video composed of 16 frames with an A800

GPU. The denoising process is executed for 25 steps with the classifier-free guidance scale set to 15. As presented in Tab. 3, we compare different efficiency strategies against the variant without any flow-related module. As previously mentioned, since our flow branch is independent of timestep during training, during the testing phase, we only need to utilize the flow branch in the first step of the denoising process to complete the damaged optical flow and cache the memory bank. For the remaining denoising steps, we can directly use the completed flow for latent interpolation and the cached K, V for flow guidance. As discussed in Sec. 4.4, we apply latent interpolation exclusively during the initial phase. Fig. 10 illustrates that continuously increasing the acceleration step beyond the early stage of the denoising process results in a sharp decline in performance. By utilizing flow-guided latent interpolation to accelerate the initial denoising and incorporating flow caching in subsequent phases, we can minimize denoising time with only a slight compromise in performance.

Thus, we determined an optimal solution: applying latent interpolation for the initial 5 steps ($2^{nd} \sim 6^{th}$) and flow caching for the remaining steps, resulting in 18.8% speedup. Compared to the pure variant, which does not incorporate flow completion and flow attention, these efficiency benefits nearly offset the additional computational burden, incurring minimal cost. Notably, FloED outperforms its closest counterpart CoCoCo [38] by 10.19s in total.

Discussion. Since no trainable modules are involved, latent interpolation can be directly adapted to diffusion-based methods like CoCoCo [38] for acceleration. We acknowledge that obtaining the completed optical flow beforehand may limit its transferability. Additionally, FloED primarily targets BR and OR tasks but is also capable of video editing. Meanwhile, compared to the conventional flow-guided method Propainter [37], FloED demonstrates superior performance in OR during flow completion. These points are further discussed in the appendix.

6. Conclusion

In this paper, we introduce FloED, an advanced video inpainting framework that efficiently incorporates flow guidance into diffusion models. By employing a dual-branch architecture, FloED first completes corrupted flow through a dedicated flow branch, followed by a
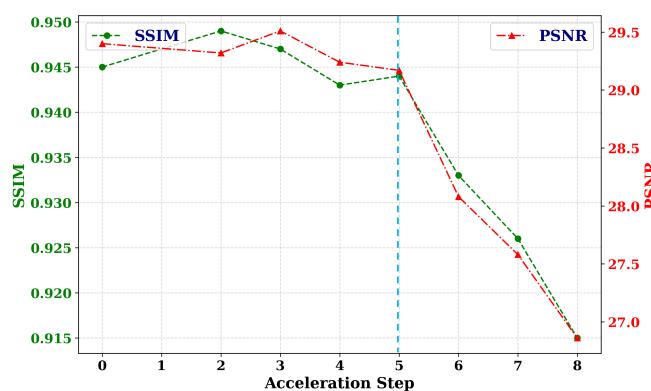

Figure 10. **Speeding steps study**. Performance markedly deteriorates when the acceleration step surpasses five.

multi-scale flow adapter that provides effective motion guidance to the primary inpainting branch. Furthermore, we developed a training-free latent interpolation technique that accelerates the multi-step denoising process. Complemented by a flow attention cache mechanism, FloED minimizes the additional computational burden introduced by incorporating optical flow. Comprehensive experiments demonstrated that FloED outperformed state-of-the-art methods in both background restoration and object removal tasks, showcasing its superior ability to maintain temporal consistency and content coherence in video inpainting.

References

[1] *Stable Diffusion v2*, 2022. https://huggingface.co/stabilityai/stable-diffusion-2-depth. 3, 5

[2] *Stable Diffusion v1.5*, 2022. https://github.com/compvis/stable-diffusion. 2, 3, 4, 5, 6

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 2

[4] Mounira Ebdelli, Olivier Le Meur, and Christine Guillemot. Video inpainting with short-term windows: Application to object removal and error concealment. *IEEE Trans. Image Process.*, 2015. 1, 2

[5] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, 2020. 1, 4

[6] Bohai Gu, Heng Fan, and Libo Zhang. Two birds, one stone: A unified framework for joint learning of image and video style transfers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23488–23497. IEEE, 2023. 7

[7] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1, 2, 3, 5

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *CoRR*, abs/2104.08718, 2021. 7

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[10] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[11] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *CoRR*, abs/2307.14073, 2023. 5

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 6

[13] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have A semantic latent space. In *ICLR*, 2023. 5, 7

[14] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 6

[15] Zhen Li, Chengze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *CVPR*, 2022. 1, 2, 4

[16] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[17] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, 2021. 1, 2

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 6

[19] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *CoRR*, 2015. 1, 2

[20] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1, 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[23] Nick C. Tang, Chiou-Ting Hsu, Chih-Wen Su, Timothy K. Shih, and Hong-Yuan Mark Liao. Video inpainting on digitized vintage films via maintaining spatiotemporal continuity. *IEEE Trans. Multim.*, 2011. 1, 2

[24] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow (extended abstract). In *IJCAI*, 2021. 4, 5, 7

[25] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Nikolai Yakovenko, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 7

[26] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *CoRR*, abs/2306.02018, 2023. 6

[27] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 2004. 7

[28] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 7

[29] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 6, 7

[30] Jinbo Xing, Menghan Xia, Yong Zhang, Hao Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, pages 399–417. Springer, 2024. 2

[31] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott Cohen, and Thomas S. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 1, 2

[32] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *CVPR*, 2019. 2

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 1, 2

[34] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020. 1, 2

[35] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *ECCV*, 2022. 2, 4

[36] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris N. Metaxas, and Licheng Yu. AVID: any-length video inpainting with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 7162–7172. IEEE, 2024. 1, 2, 3, 7

[37] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. *CoRR*, abs/2309.03897, 2023. 1, 2, 4, 6, 8

[38] Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Kam-Fai Wong, and Lei Zhang. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. *CoRR*, abs/2403.12035, 2024. 1, 2, 3, 6, 8