Keyframe-Guided Creative Video Inpainting

Yuwei Guo[1] Ceyuan Yang[4] Anyi Rao[7] Chenlin Meng[3] Omer Bar-Tal[3]
Shuangrui Ding[1] Maneesh Agrawala[6] Dahua Lin[1,2,5] Bo Dai[8,9]

[1]CUHK [2]Shanghai AI Laboratory [3]Pika Labs [4]ByteDance [5]CPII under InnoHK
[6]Stanford [7]HKUST [8]HKU [9]Feeling AI

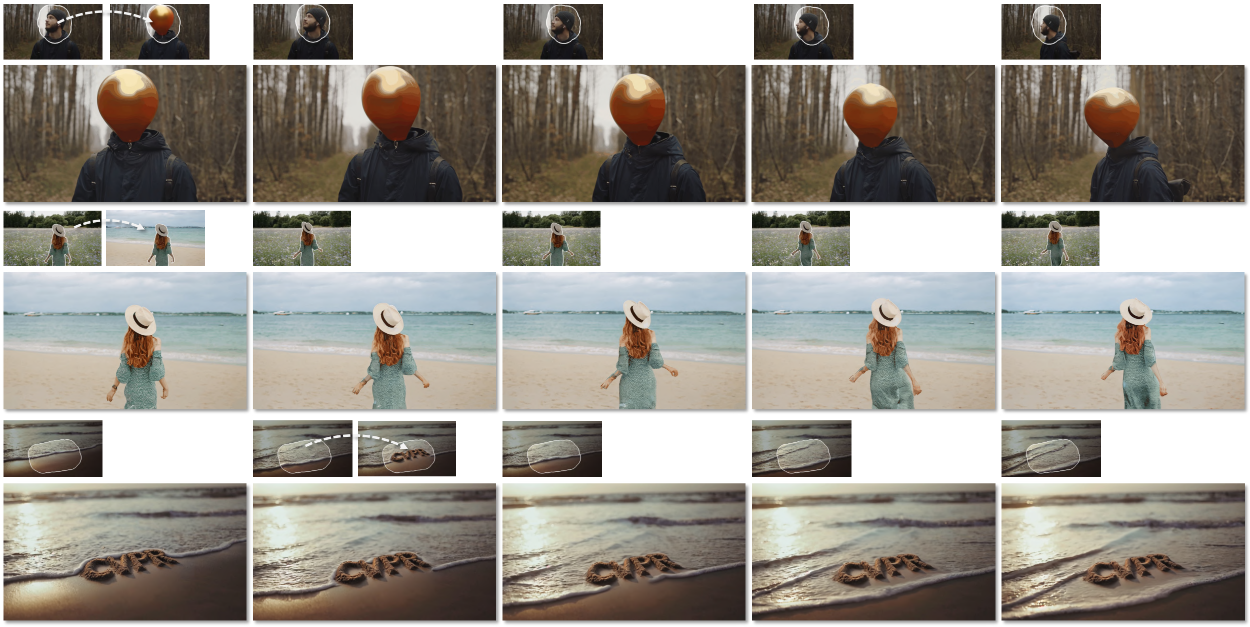


Figure 1. We introduce VideoRepainter, a practical framework for video inpainting using a keyframe as reference. We show example use cases such as changing a person's head into a balloon (object changing, first row), altering the environment from field to beach (background changing, second row), and adding a sand-made "CVPR" to the beach (novel concept insertion, third row).

Abstract

Video inpainting, which aims to fill missing regions with visually coherent content, has emerged as a crucial technique for creative applications such as editing. While existing approaches achieve visual consistency or text-guided generation, they often struggle to balance coherence and creative diversity. In this work, we introduce VideoRepainter, a two-stage framework that allows users to inpaint a keyframe using established image-level techniques, then propagate the changes to other frames. Our approach can leverage state-of-the-art image models for keyframe manipulation, thereby easing the burden of the video-inpainting process. To this end, we integrate an image-to-video model with a symmetric condition mechanism to address ambiguity caused by direct mask downsampling. We further explore efficient strategies for mask synthesis and parameter tuning to reduce costs in data processing and model training. Evaluations demonstrate our method achieves superior results in both visual fidelity and content diversity compared to existing approaches, providing a practical solution for creative video manipulation. See our Project Page for more details.

1. Introduction

Video inpainting aims to fill the missing regions of a video with coherent content. Earlier efforts in this domain focused primarily on restoration tasks, where the objective was to repair missing content by ensuring harmony with the surrounding visual context. However, the emergence of advanced video generation models (*e.g.*, Pika [51], Kling [34], Gen3 [57]) has transformed user expectations beyond mere reconstruction. Contemporary applications demand creative content manipulation, enabling users to modify existing videos in innovative ways. This includes incorporating novel objects or content through instruction-guided inpainting, as well as environment modifications via background replacement for virtual tour applications. These evolving requirements present new challenges in achieving visual coherence with creative flexibility, necessitating novel approaches to video inpainting.

While maintaining visual coherence remains crucial, a significant challenge in creative video inpainting lies in achieving user control over the regenerated regions. Current approaches often utilize pre-trained text-to-video (T2V) diffusion models, employing textual prompts as inpainting guidance. Recent works such as AVID [89] and CoCoCo [93] have attempted to enhance this approach by fine-tuning T2V models with masked pixel conditioning. However, these methods, which rely heavily on pre-trained T2V backbones, frequently encounter limitations in visual quality and controllability. These limitations stem primarily from the relative scarcity of video training data, resulting in inferior performance (*e.g.*, weak text-following ability) of T2V models relative to their text-to-image counterparts. Furthermore, the slow video generation processes compound these issues, leading to a cumbersome user experience that often requires extensive trial and error to achieve desired results.

In this work, we provide VideoRepainter for the task of video inpainting by disentangling the generation of still content and dynamic motion. Specifically, VideoRepainter first allows users to inpaint a keyframe using any established image-level techniques [7, 28, 30] and then propagates these modifications across the temporal dimension. This decoupled approach can leverage the superior visual quality and controllability of state-of-the-art image inpainting methods, thereby expanding the possibilities for user-directed content manipulation. Furthermore, by isolating temporal aspects, the video generation component can focus specifically on maintaining coherence across frames rather than content generation. To address the critical challenge of temporal consistency, we exploit the similarity between image-to-video (I2V) generation and video inpainting, *i.e.*, both being video generation tasks conditioned on partially observed pixels. Our investigation reveals that conventional downsampled mask conditioning introduces ambiguity in video inpainting, which we resolve through a symmetric mask conditioning approach. To support arbitrary user-defined masks and facilitate training on in-the-wild videos, we implement efficient mask synthesis strategies. Additionally, we demonstrate that our I2V model repurposing can be achieved by updating less than 2% of the model parameters, significantly enhancing training efficiency.

Extensive experiments demonstrate that VideoRepainter achieves superior visual quality and versatility in video inpainting tasks while maintaining modest computational and data requirements in training. By leveraging state-of-the-art image inpainting capabilities, our approach enables sophisticated visual effects challenging for prior text-guided methods, such as transforming a person's head into a ballon or adding a sand-made "CVPR" on the beach (Fig. 1). Evaluations across diverse use cases consistently show that our method outperforms existing approaches in both perceptual quality and content diversity, validating its effectiveness as a practical solution for creative video manipulation.

2. Related Work

Image Inpainting. Image inpainting has been a long-standing problem in computer vision for years. Earlier explorations in this field primarily involve traditional approaches [6, 14], Variational Autoencoders (VAEs) [49, 64, 91], and Generative Adverserial Networks (GANs) [15, 39, 48, 58, 63, 70, 80, 81, 84]. With the development of Diffusion Models (DMs) [25, 55, 62], recent focus has shifted to leverage DMs to solve the inpainting problem. One branch of this work leverages the DMs' zero-shot inpainting capabilities [2, 3, 13, 38, 79, 85]. Repaint [44] alters the denoising by sampling the unmasked regions using the given image information. Other work enables inpainting by training dedicated models with masked pixels as conditioning [55, 69, 73, 74, 78, 83, 85]. For instance, Stable Diffusion Inapint [55] directly takes the mask and masked image as input to the UNet [56] and finetunes a text-to-image diffusion model to predict the complete image. SmartBrush [73] incorporates shape guidance in addition to a text prompt for more flexible content control. BrushNet [30] introduces a plug-and-play UNet branch to separately tackle the image features and noisy latent to achieve flexible inpainting. Our method can leverage state-of-the-art image inpainting to the context of video inpainting.

Video Inpainting. Traditional video inpainting focuses on restoring the missing area of a video. Earlier work in this field employs convolutional networks for spatiotemporal information aggregation [12, 27, 67]. Other work leverages optical flow as additional information [22, 33, 36, 37, 77, 94], or video Transformer [9, 40, 41, 92] for video inpainting. However, these methods primarily focus on content coherence instead of controllability. Recently, another line of research has tackled the problem of text-guided video inpainting by leveraging the priors of video diffusion models. For instance, AVID [89] adopts a pre-trained video motion module [24] and trains the inpainting model with structure guidance, enabling the task of inpainting-based video editing with a text prompt. CoCoCo [93] introduces an enhanced motion capture module for better inpainting consistency and compatibility with multiple backbone variants. Though these methods can inpaint with text-aligned guidance, they sometimes suffer from low visual quality tied to the base video diffusion models. We tackle the video inpainting by first solving a keyframe inpainting problem, thereby pushing the inpainting quality to a higher level.

Video Editing. Video editing aims to change a certain aspect of a source video, *e.g.*, global style, local region, content movement [65, 90], *etc*. This area has seen significant recent progress with the development of image and video diffusion models [4, 11, 16, 20, 21, 29, 31, 42, 43, 45, 46, 61, 71, 88]. Some of these methods leverage a pre-trained text-to-image prior and enforce temporal consistency via feature fusion [10, 23, 52]. Other methods explore various approaches such as one-shot tuning [47, 72] and neural atlases [4, 11], *etc*. Our works feature re-inpainting a source video within a specified region and can enable localized video editing with higher visual quality and diversity.

3. Methods

Given a video sequence $\{f^i\}_{i=1}^N$ with corresponding mask sequence $\{m^i\}_{i=1}^N$ (derived from user input or segmentation models [54]) and an optionally inpainted keyframe f^j_{edit}, we aim to generate visually coherent content for masked regions while maintaining consistency with known areas.

We present the key components of VideoRepainter: Sec. 3.1 details the adaptation of image-to-video diffusion models for inpainting with resolved mask ambiguity; Sec. 3.2 introduces our robust training mask generation strategy for in-the-wild videos; and Sec. 3.3 presents the training and inference implementations.

3.1. Efficient Model Repurposing

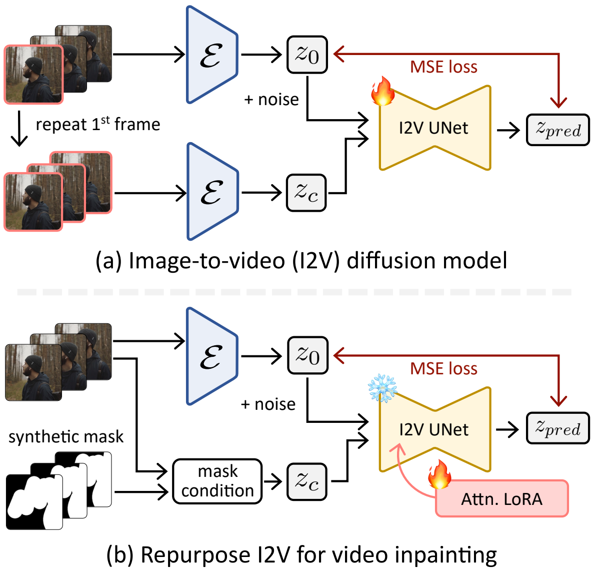
Video inpainting can be formulated as a conditional generation task, where the model synthesizes content based on partially observed pixels. This formulation closely aligns with image-to-video (I2V) generation, where synthesis is conditioned on the initial frame. Given this structural similarity and shared underlying requirements, we opt to leverage pretrain I2V priors for video inpainting through efficient model repurposing, substantially reducing computational and data requirements compared to training from scratch (Fig. 2).

3.1.1 Image-to-Video Diffusion Model

Image-to-video (I2V) models perform the video generation task conditioned on the first frame. Representative I2V models such as Stable Video Diffusion (SVD) [8] and DynamiCrafter [75] are built upon powerful text-to-image diffusion models [55]. They expand the T2I model by adding temporal layers, *e.g.*, temporal convolution and attention, to model cross-frame consistency and motion dynamics. As the additional condition, the first frame latents z_0^0 encoded by the VAE are temporally repeated and concatenated to the noisy sample z_t^i to form the denoising UNet's input $z_{t,c} = [z_t^i, z_0^0]$, as shown in Fig. 2 (a). Here z_j^i represents the i-th frame's latent at j-th diffusion timesteps. Moreover, the CLIP [53] text embedding is replaced by the first frame's embedding computed from the CLIP's vision branch. The denoising UNet is trained under the EDM [32] framework, learning to predict a cleaner version of the noisy sample:

$$\mathcal{L} = \mathbb{E}_{z_0,t,\epsilon\sim\mathbb{N}(0,\sigma^2)}[\|z_0 - z_\theta(z_{t,c},t,c)\|_2^2], \tag{1}$$

where c is the conditional vision embedding, z_t is noised sample, t is the diffusion timestep condition. In this work, we adopt SVD [8] as the I2V backbone.



(a) Image-to-video (I2V) diffusion model

(b) Repurpose I2V for video inpainting

Figure 2. **Image-to-Video (I2V) Model Repurposing.** (a) I2V diffusion models are conditioned on repeated first frame latent sequence obtained from VAE encoder $\mathcal{E}(\cdot)$; (b) We efficiently repurpose I2V for video inpainting by replacing z_c with our symmetric mask condition (see Sec. 3.1.2) and LoRA [26] finetuning. Here we visualize with z_0-prediction for clarity.

3.1.2 Resolving Mask Ambiguity

The video inpainting model requires both pixel values of known regions and precise mask information. While image-to-video tasks operate with fixed uncertainty regions (all frames except the first), inpainting tasks must handle arbitrary masked regions. Thus, additional information should be considered as the model's input.

Conventional mask conditioning approaches typically concatenate binary masks with UNet inputs [30, 55, 89, 93] (Fig. 3 (a)). However, such practice introduces ambiguity for methods based on latent diffusion models

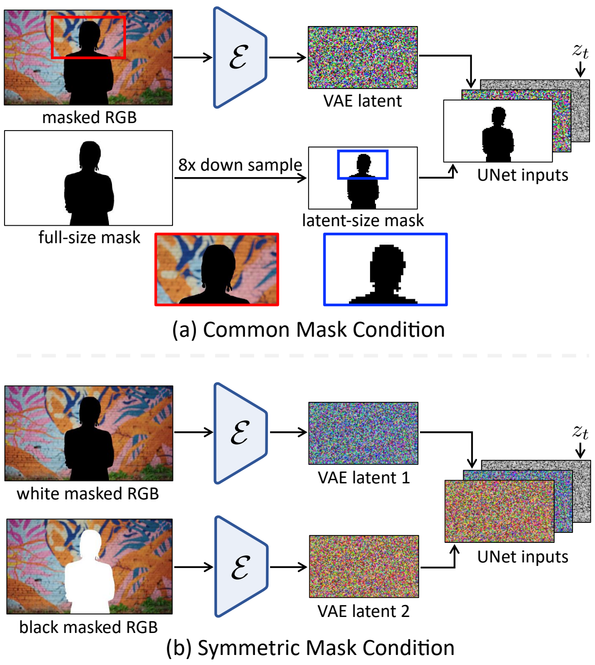


Figure 3. **Comparing Mask Conditioning Mechanism.** (a) Conventional diffusion-based inpainting approaches [30, 55] downsample masks to match VAE latent dimensions, compromising fine-grained boundary details. (b) Our proposed symmetric mask conditioning encodes dual variants of the masked image to preserve mask fidelity and resolve spatial ambiguity.

(LDMs) [55]. LDMs operate in a compact latent space with downsampled spatial dimensions (*e.g.*, the downsampling ratio is 8 for Stable Diffusion), and thus require corresponding mask downsampling to match the latent size. This resolution reduction compromises mask precision, particularly evident in the degradation from detailed RGB masks to jagged representations, as shown in the red and blue crop in Fig. 3 (a). Consequently, pixels within complex mask boundaries may be misclassified into unmasked regions, as demonstrated by the earring shape in our example. Our experiments reveal that this issue becomes particularly problematic when handling complex dynamic masks (Sec. 4.3).

We address this ambiguity through symmetric mask encoding (Fig. 3 (b)). Our approach encodes two variants of the masked image through the VAE, one with black-filled mask regions and another with white-filled regions, and concatenates both encoded latents as UNet conditions. This enables precise pixel classification: identical values across variants indicate unmasked regions, while differences denote masked areas. Moreover, this design maintains compatibility with the VAE latent space and thus facilitates model repurposing. In implementations, we newly involve a trainable input convolution layer to the UNet to process the additional channels (Fig. 2 (b)).

3.1.3 Reusing Pre-trained Knowledge

Both image-to-video generation and video inpainting tasks operate as pixel-conditioned generation, as noted in Sec. 3.1. To maximize the utility of pre-trained knowledge, we minimize network modifications by optimizing a small subset of parameters. We adopt Low-Rank Adaptation (LoRA) [26] for efficient task adaptation. LoRA approximates weight updates through two low-rank matrices: $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{n\times r}$, transforming pre-trained weights W to $W' = W + AB^T$. These trainable rank-decomposition matrices enable efficient parameter updates while preserving the core model structure.

In our implementation, we apply such a technique to spatial and temporal attention layers, as these components are crucial for capturing long-range visual correspondences. This adaptation strategy proves highly effective, with less than 2% of the parameters being updated compared to the entire model finetuning. This minimal modification significantly reduces memory usage, making our method more accessible for practical applications.

3.2. Supporting Arbitrary Inpainting Region

Our goal is to develop a unified model supporting diverse inpainting applications, including keyframe-based inpainting, background modification, object insertion, *etc*. This requires robust handling of spatially and temporally arbitrary inpainting regions. The model must process various mask types, ranging from precise video object segmentation to simple rectangular regions and rough user sketches.

Previous approaches have relied on online instance detection [93] or video segmentation datasets [76, 92]. However, these methods introduce computational overhead or data preprocessing requirements. We demonstrate that applying robust content-agnostic mask augmentation strategies during training is enough to achieve strong generalization capabilities. This approach enables efficient scaling to in-the-wild videos without preprocessing overhead. Our mask augmentation strategy incorporates the following techniques:

Spatial augmentations. For the spatial mask shape, we adopt the following mask types and transformations:

- *Full mask.* All pixels are masked, aiming to maintain the model's generation ability;
- *Grid mask.* We split spatial regions into grids and mask each grid independently under a fixed probability.
- *Square mask.* A random square region is masked.
- *Scribble mask.* Random scribble are generated to stimulate human drawing behaviours [82].
- *Mask reversion.* All above mask types except that the full mask is randomly reversed to enhance applications such as changing the background.

Temporal augmentations. For each training video, we first perform spatial augmentations to obtain an initial mask. We then apply the following temporal transformations to get a mask sequence for the video.

- *Static.* The initial mask is copied to all frames.
- *Spatiotemporal bezier.* A random Bezier curve is generated to serve as the moving trajectory of the mask. The initial mask then moves along the trajectory to simulate masks that track specific objects.
- *Temporal variants.* After obtaining the mask sequence, some shape randomness is randomly added to each frame to stimulate the mask shape change over time.

Task-specific augmentation. We randomly make some frames fully visible to the model to support inpainting with keyframe reference. In other cases, the model is optimized for unconditioned inpainting.

Since none of the above augmentations involve deep learning models, they can be efficiently generated on the fly without placing a noticeable burden on training time.

3.3. Implementations

Training. Fig. 2 (b) illustrates our training procedure. Training video clips undergo random masking (Sec. 3.2), followed by symmetric VAE encoding to obtain the mask condition z_c (Fig. 3), which is then concatenated with the noisy sample z_t as UNet input. While maintaining the objective function in Eq. (1), we update only the input convolution and LoRA parameters. This parameter-efficient approach, which preserves most pre-trained weights, effectively prevents overfitting. We thus employ a progressive training strategy by initially efficient training at a lower resolution, followed by high-resolution fine-tuning.

Extension Beyond Training Length. Video diffusion models typically operate on fixed-length sequences, with direct inference on longer sequences often resulting in quality degradation. While existing approaches employ temporal MultiDiffusion [5], *i.e.*, denoising overlapped clips simultaneously with averaged overlapping regions, it lacks consistency constraints and may suffer from content drifting.

We address this limitation through a coarse-to-fine generation process for sequences exceeding the training length. First, we sample sparse frames uniformly across the video duration for initial content propagation. These inpainted frames then serve as anchors for the second stage, where we apply MultiDiffusion with each subsequence conditioned on the anchor frames, ensuring fidelity to the original keyframe while generating the complete sequence.

4. Experiments

We choose Stable Video Diffusion (SVD) [8] as our image-to-video (I2V) backbone. It is able to generate 14 (`svd-base`) or 25 (`svd-xt`) frames. We chose `svd-xt` for the main experiments. After initialization, the UNet contains 1525.9 M parameters and 27.9 M trainable parameters. We train the model with a self-collected watermark-free video dataset that contains about 300K videos. The low-resolution pre-training is conducted on 320×576, and high-resolution fine-tuning is on 576×1024. We use AdamW optimizer and set the learning rate to $5e-5$. We use 16 NVIDIA A100s for training. The total optimization iteration is 50K.

We infer the model with Euler sampler [32] in the Diffuser library [66] with 25 sampling steps. Other hyperparameters are fixed to the original SVD settings. For the first stage of image inpainting, we use Adobe Photoshop [28] generative infilling for background replacement usage and FLUX [7] inpainting ControlNet [1, 86] for other cases.

4.1. Qualitative Results

Our method achieves exceptional visual quality and diversity in creative video inpainting, leveraging state-of-the-art image inpainting techniques [1, 28] and video generation priors from the I2V model [8]. As demonstrated in Fig. 1 and Fig. 4, our method successfully addresses various creative inpainting scenarios, including novel object addition, background modification, virtual try-on, *etc*. Our approach enables sophisticated visual manipulations that were previously challenging for existing methods, such as transforming a human head into a balloon (Fig. 1 case 1), integrating sand-textured text into beach scenes (Fig. 1 case 3), compositing tiny villages within coffee cups (Fig. 4 case 2), and placing sailboats among clouds (Fig. 4 case 3). This creative capability stems from our two-stage inpainting architecture combined with advanced image inpainting models. The I2V model's temporal priors ensure seamless integration of inpainted content, maintaining both visual and temporal coherence with unmasked regions, as evidenced by consistent background perspective changes (Fig. 1 case 2, Fig. 4 case 5) and natural water-sand interactions (Fig. 1 case 3).

4.2. Comparisons to Prior Works

Baselines. We consider the following baseline covering recent progress in video inpainting, generation, and editing: (1) *CoCoCo* [93]: a text-guided video inpainting framework featuring motion capture modules for temporal consistency. While AVID [89] shares similar architectural designs, its implementation remains close-sourced at the time of our experiments. Therefore, we adopt CoCoCo as our primary text-guided baseline, given its comparable technical approach and public accessibility. (2) *I2VGen-XL* [87]: an image-to-video diffusion model trained on high-quality and high-resolution videos. (3) *ModelScope* [68]: a text-to-video diffusion model trained on large-scale text-video paired data. We adapt these models for video inpainting following previous practices [2, 44, 89]; (4) *AnyV2V* [35]:

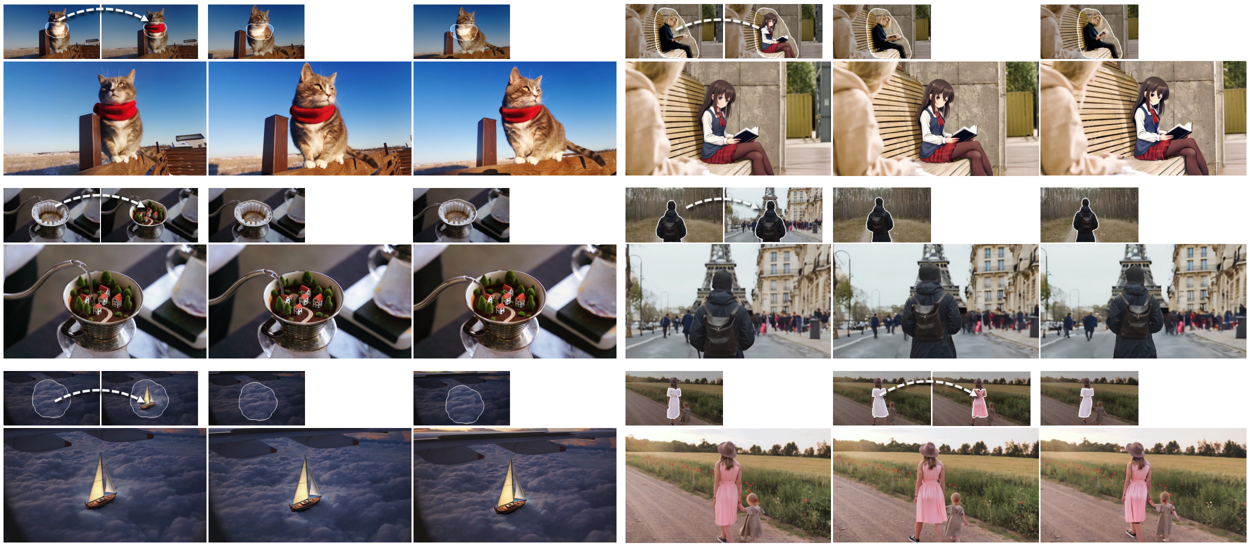


Figure 4. **Qualitative Results.** Here we demonstrate some creative use cases, *e.g.*, novel object insertion, background changing, and virtual try-on. Our method propagates the keyframe inpainting results throughout the whole video with excellent consistency and visual quality. *We recommend referring to our Project Page for better visualizations.*

Metrics	Cons.$_{\uparrow}$	PSNR (Bg.)$_{\uparrow}$	MSE (Bg.)$_{\downarrow\times 10^4}$	AS$_{\uparrow}$
CoCoCo [93]	93.94	28.33	18.82	**4.57**
I2VGen-XL [87]	93.52	29.51	13.09	4.34
ModelScope [68]	91.55	25.18	34.16	2.30
Ours	**96.01**	**31.46**	**8.85**	4.47

Table 1. **Video Inpainting Evaluations.** “Cons.” stands for CLIP consistency score; “AS” stands for aesthetic score; “Bg.” stands for background fidelity metrics.

Metrics	Cons.$_{\uparrow}$	PSNR (Bg.)$_{\uparrow}$	MSE (Bg.)$_{\downarrow\times 10^4}$	AS$_{\uparrow}$
CoCoCo [93]	96.16	31.82	6.66	5.24
I2VGen-XL [87]	95.98	31.99	7.61	4.68
ModelScope [68]	95.32	25.77	28.28	3.77
AnyV2V [35]	96.28	18.38	173.96	5.44
Ours	**97.31**	**32.29**	**3.77**	**5.49**

Table 2. **Video Editing Evaluations.** Abbreviations remain the same with Tab. 1.

a state-of-the-art video editing framework that employs DDIM inversion [17, 62] to preserve source video characteristics while enabling spatiotemporal feature injection.

Settings. We compare our method against baselines on video inpainting and editing tasks. For video inpainting, we omit the editing method AnyV2V since it’s out of the baseline’s scope. For video editing, we provide an inpainted first frame for I2VGen-XL and AnyV2V, while providing detailed text prompts for CoCoCo and ModelScope since they do not accept image conditioning.

4.2.1 Quantitative Comparisons

We quantitatively compare our method against baselines on video inpainting and inpainting-based editing. For inpainting evaluation, our model runs in unconditioned inpainting mode and does not receive a keyframe. We consider the following dimensions and derive the corresponding metrics that align with prior work: (1) *Temporal consistency*: the smoothness of the final video, measured by the CLIP vision embedding’s distance of all the adjacent frames [24, 53, 72, 89]; (2) *Background fidelity*: following [30], we evaluate unmasked regions’ fidelity to the source video with PSNR and MSE error; (3) *Visual quality*: the perframe aesthetics score [30, 59, 60], revealing the leval of artifacts and visual harmony. For inpainting evaluation, we use 50 videos from the DAVIS dataset [50], and for the editing evaluation, we create 20 editing samples based on in-the-wild videos.

As shown in Tab. 1 and Tab. 2, our method outperforms the others in most aspects. Compared to the prior editing state-of-the-art method AnyV2V, our method provides better background fidelity in an end-to-end pipeline design, without the need for DDIM inversion and regeneration process, thereby significantly saving the inference compute.

4.2.2 Qualitative Comparions

We present comparative editing results in Fig. 5, focusing on two challenging scenarios: background alteration and novel object insertion. Text-based inpainting methods CoCoCo and ModelScope generate semantically appropriate

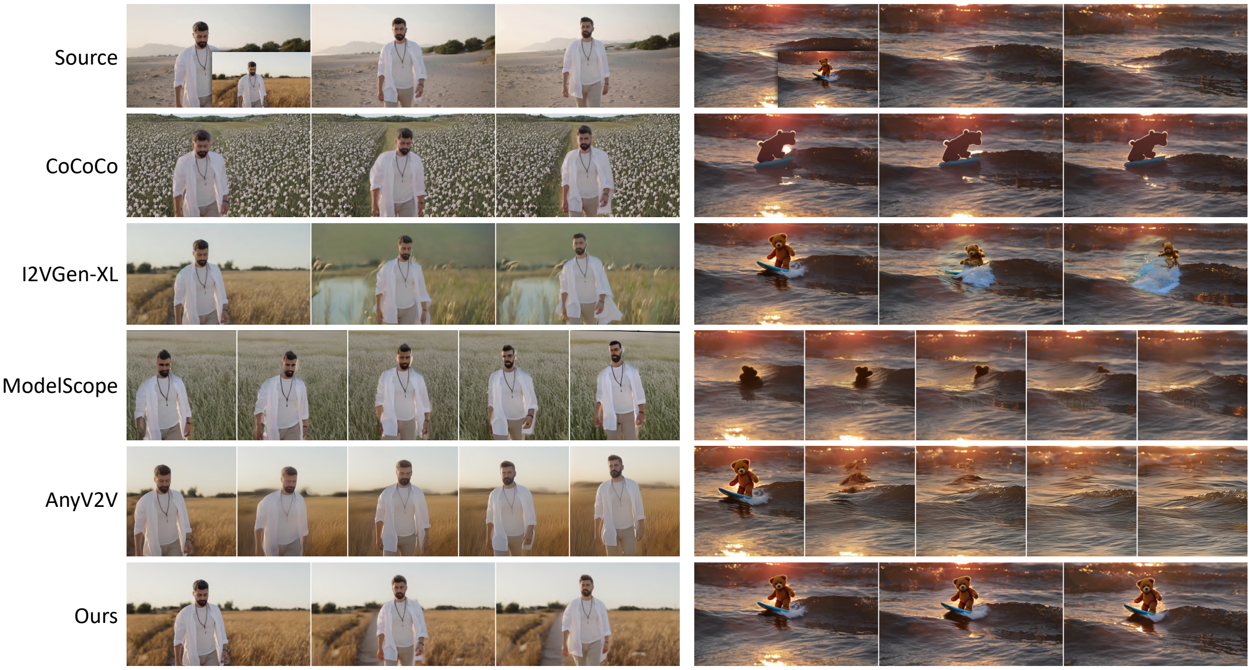


Figure 5. **Qualitative Comparisons.** We compare VideoRepainter with representative text-guided inpainting models, text-/image-to-video generation models, and keyframe-based video editing approach. On challenging use cases such as background changing and new object insertions, our method produces preferred visual quality and consistency.

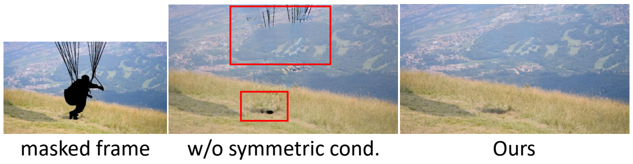


Figure 6. **Symmetric Mask Condition.** Using downsampled mask (Fig. 3 (a)) leads to ambiguity and the leakage of black pixels, and our solution resolves such an issue.

Metrics	#Param	Cons.$_{\uparrow}$	PSNR (Bg.)$_{\uparrow}$	MSE (Bg.)$_{\downarrow \times 10^4}$	AS$_{\uparrow}$
w/o symm.	27.91 M	95.17	27.52	22.23	4.20
w/o full mask	27.91 M	96.46	28.28	18.27	4.21
w/o grid mask	27.91 M	96.48	28.56	18.24	3.97
w/o T-bezier	27.91 M	94.87	30.08	12.50	4.19
w/o T-variants	27.91 M	96.52	29.65	13.42	3.94
full model	1525.94 M	97.48	28.90	17.11	4.26
only attn.	771.78 M	96.71	30.80	8.90	4.32
Ours	27.91 M	96.73	30.78	11.59	4.24

Table 3. **Ablative Comparisons.** We ablate the symmetric mask condition, mask synthesis strategy, and trainable components by training several model variants on the `svd-base` backbone.

content but exhibit limitations in visual quality, particularly evident in large-region modifications and uncommon semantic elements like the teddy bear case. I2VGen-XL demonstrates detail loss and temporal inconsistencies during frame propagation. While AnyV2V maintains background consistency, it suffers from reduced high-frequency detail. Additionally, its DDIM inversion approach lacks motion cues for newly introduced objects, resulting in the disappearance of the teddy bear into the ocean waves. Our method achieves superior visual consistency with the edited frame while maintaining stable, high-quality output across sequences, as demonstrated in the bottom row of Fig. 5.

4.3. Ablating Design Components

We conduct ablations on some essential design components of our method. To maintain efficiency and an economical compute budget, our ablation models are trained with the `svd-base` backbone and a total number of 14 frames.

Symmetric Mask Condition. We evaluate our symmetric condition mechanism against the conventional downsampled mask approach illustrated in Fig. 3. Our symmetric condition preserves mask detail fidelity, particularly for complex mask geometries. The comparison in Fig. 6 demonstrates that downsampled masks can lead to misclassification of masked pixels as unmasked regions, producing black pixel artifacts in complex boundary areas (case 2). In contrast, our method achieves smooth transitions between inpainted and original regions (case 3). Quantitative evaluation further supports these observations. As shown in Tab. 3, comparing “w/o symm.” versus “Ours” reveals

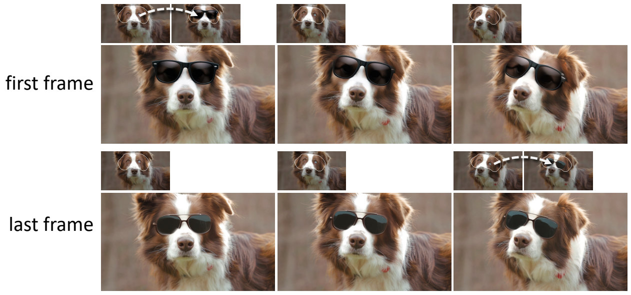


Figure 7. **Keyframe Location.** Our method consistently produces high-fidelity results regardless of keyframe location.

that removing the symmetric mask condition reduces background fidelity.

Mask Synthesis Strategy. We evaluate our masking strategies designed for uncurated Internet video training and arbitrary mask inference through controlled experiments reported in Tab. 3. Results indicate that removing spatial mask augmentation reduces background fidelity scores while eliminating temporal augmentation decreases temporal consistency metrics. The combination of all masking strategies in our final model achieves optimal performance across evaluation metrics, demonstrating the complementary nature of these techniques.

Trainable Components. We evaluate the efficacy of our adapter training strategy through comparative analysis in Tab. 3, examining three variants: full model parameter training ("full model"), attention-only training ("only attn."), and only attention LoRA adapters ("Ours"). While unrestricted model optimization offers greater flexibility, it leads to decreased background fidelity, likely due to overfitting on limited training data. Attention layer tuning yields better results through parameter regularization, with our adapter approach achieving comparable performance while maintaining optimal parameter efficiency.

Keyframe Location. While using the first frame as the inpainting keyframe is conventional, supporting arbitrary keyframe selection provides crucial flexibility when initial frames contain challenging occlusions or deformations. Despite utilizing an image-to-video backbone, our method maintains consistent performance regardless of keyframe position. As demonstrated in Fig. 7, our approach achieves high-fidelity inpainting results with both first-frame and last-frame keyframe configurations, highlighting its temporal robustness.

Inference Mask Precision. When training the model, we use random and content-agnostic masks, which not only offer efficient synthesis but also prevent the model from learning unwanted correlations between objects and mask shapes or positions. This design choice also offers flexibility in inference by supporting arbitrary mask precision levels, as shown in Fig. 8. Users can either leverage video segmentation models [18, 19, 54] for a precisely localized editing, or arbitrarily draw the region for a coarse indication.

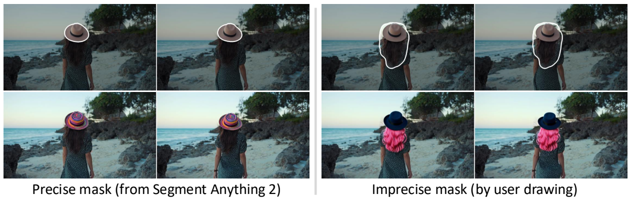

Precise mask (from Segment Anything 2)
Imprecise mask (by user drawing)

Figure 8. **Mask Precision.** Our model generalizes across varying levels of mask precision and offers more editing flexibility.

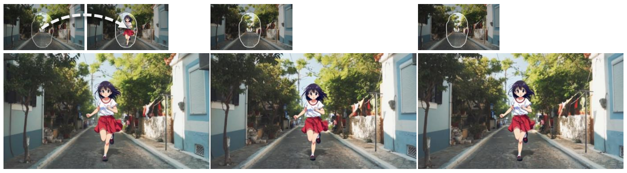


Figure 9. **Failure Case.** Our method exhibits little or no motion when the inpainting is beyond the domain capacity of the underlying I2V model, *e.g.*, adding an anime girl to a realistic street.

5. Conclusion

We present VideoRepainter, a practical framework that advances creative video inpainting by enhanced content diversity and visual fidelity. Our approach introduces a two-stage architecture that first employs established image inpainting models for keyframe modification, followed by temporal propagation to the remaining frames. The framework features an efficient adaptation of pre-trained image-to-video models, incorporating symmetric mask conditioning and spatiotemporal augmentations to ensure precise mask handling and data-efficient training. Extensive evaluations across diverse scenarios demonstrate that VideoRepainter consistently outperforms existing methods in both perceptual quality and creative expression.

Limitations. The effectiveness of our approach is inherently bounded by both the underlying image-to-video (I2V) backbone and the selected image inpainting model. Specifically, we find a common failure mode where the keyframe inpainting is beyond the domain capacity of the underlying I2V model, as shown in Fig. 9. In such cases, the newly introduced content typically exhibits little or no motion. Moreover, we assume having a user-provided mask region, but instructional-based editing is preferred in authentic use cases. Future improvements may address these constraints through enhanced I2V architectures and diverse temporal training data, as well as involving feedback from Multi-modal Large Language Models (MLLMs).

Acknowledgements

This project is partially supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, HKU Startup Fund, and HKU Shanghai Intelligent Computing Research Center. We thank the anonymous reviewers for their comments that helped improve this paper.

References

[1] alimama creative. Flux-controlnet-inpainting, 2024. 5

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 2, 5

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42 (4):1–11, 2023. 2

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 3

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 5

[6] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[7] black-forest labs. flux, 2024. 2, 5

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 5

[9] Jiayin Cai, Changlin Li, Xin Tao, Chun Yuan, and Yu-Wing Tai. Devit: Deformed vision transformers in video inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 779–789, 2022. 2

[10] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 3

[11] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 3

[12] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9066–9075, 2019. 2

[13] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024. 2

[14] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9): 1200–1212, 2004. 2

[15] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018. 2

[16] Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. In *European Conference on Computer Vision*, pages 183–199. Springer, 2025. 3

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6

[18] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024. 8

[19] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. In *European Conference on Computer Vision*, pages 215–233. Springer, 2024. 8

[20] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3

[21] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6712–6722, 2024. 3

[22] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 713–729. Springer, 2020. 2

[23] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[24] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 6

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 4

[27] Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. Proposal-based video completion. In *Computer Vision–ECCV 2020: 16th European*

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, pages 38–54. Springer, 2020. 2

[28] Adobe Inc. Adobe photoshop, 2023. 2, 5

[29] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023. 3

[30] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 2, 3, 4, 6

[31] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024. 3

[32] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3, 5

[33] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5792–5801, 2019. 2

[34] Kling AI. https://klingai.com/, 2024. 1

[35] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 5, 6

[36] Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 728–743. Springer, 2020. 2

[37] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17562–17571, 2022. 2

[38] Anji Liu, Mathias Niepert, and Guy Van den Broeck. Image inpainting via tractable steering of diffusion models. *arXiv preprint arXiv:2401.03349*, 2023. 2

[39] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9371–9381, 2021. 2

[40] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021. 2

[41] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14040–14049, 2021. 2

[42] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 3

[43] Tianyi Lu, Xing Zhang, Jiaxi Gu, Renjing Pei, Songcen Xu, Xingjun Ma, Hang Xu, and Zuxuan Wu. Fuse your latents: Video editing with multi-source latent diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6745–6754, 2024. 3

[44] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2, 5

[45] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3

[46] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *arXiv preprint arXiv:2405.13865*, 2024. 3

[47] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *arXiv preprint arXiv:2405.16537*, 2024. 3

[48] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[49] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10775–10784, 2021. 2

[50] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 6

[51] Pika Labs. https://pika.art/, 2024. 1

[52] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 3

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[54] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman

Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 8

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[57] RunwayML. https://academy.runwayml.com/gen3-alpha, 2024. 2

[58] Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7335–7345, 2023. 2

[59] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6

[60] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[61] Uriel Singer, Amit Zohar, Yuval Kirstain, Shelly Sheynin, Adam Polyak, Devi Parikh, and Yaniv Taigman. Video editing via factorized diffusion distillation. *arXiv preprint arXiv:2403.09334*, 2024. 3

[62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 6

[63] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[64] Ching-Ting Tu and Yi-Fu Chen. Facial image inpainting with variational autoencoder. In *2019 2nd international conference of intelligent robotic and control engineering (IRCE)*, pages 119–122. IEEE, 2019. 2

[65] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2024. 3

[66] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[67] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5232–5239, 2019. 2

[68] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 5, 6

[69] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023. 2

[70] Wentao Wang, Li Niu, Jianfu Zhang, Xue Yang, and Liqing Zhang. Dual-path image inpainting with auxiliary gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11421–11430, 2022. 2

[71] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[72] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3, 6

[73] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 2

[74] Shaoan Xie, Yang Zhao, Zhisheng Xiao, Kelvin CK Chan, Yandong Li, Yanwu Xu, Kun Zhang, and Tingbo Hou. Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. *arXiv preprint arXiv:2312.03771*, 2023. 2

[75] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin 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 *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3

[76] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 4

[77] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3723–3732, 2019. 2

[78] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3190–3199, 2023. 2

[79] Siyuan Yang, Lu Zhang, Liqian Ma, Yu Liu, JingJing Fu, and You He. Magicremover: Tuning-free text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2310.02848*, 2023. 2

[80] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 2

[81] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[82] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 4

[83] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2

[84] Yongsheng Yu, Libo Zhang, Heng Fan, and Tiejian Luo. High-fidelity image inpainting with gan inversion. In *European Conference on Computer Vision*, pages 242–258. Springer, 2022. 2

[85] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. *arXiv e-prints*, pages arXiv–2304, 2023. 2

[86] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5

[87] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 5, 6

[88] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[89] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2024. 2, 3, 5, 6

[90] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2025. 3

[91] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2

[92] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10477–10486, 2023. 2, 4

[93] 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. *arXiv preprint arXiv:2403.12035*, 2024. 2, 3, 4, 5, 6

[94] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16448–16457, 2021. 2