# Temporally Consistent Object Editing in Videos using Extended Attention

AmirHossein Zamani[1,2] * Amir G. Aghdam[1] Tiberiu Popa[1] Eugene Belilovsky [1,2]
[1] Concordia University, Montreal, QC, Canada; [2] Mila – Quebec AI Institute

## Abstract

*Image generation and editing have seen a great deal of advancements with the rise of large-scale diffusion models that allow user control of different modalities such as text, mask, depth maps, etc. However, controlled editing of videos still lags behind. Prior work in this area has focused on using 2D diffusion models to globally change the style of an existing video. On the other hand, in many practical applications, editing localized parts of the video is critical. In this work, we propose a method to edit videos using a pre-trained inpainting image diffusion model. We systematically redesign the forward path of the model by replacing the self-attention modules with an extended version of attention modules that creates frame-level dependencies. In this way, we ensure that the edited information will be consistent across all the video frames no matter what the shape and position of the masked area is. We qualitatively compare our results with state-of-the-art in terms of accuracy on several video editing tasks like object retargeting, object replacement, and object removal tasks. Simulations demonstrate the superior performance of the proposed strategy.*

## 1. Introduction

In recent years, significant strides have been made in large-scale generative modeling with diffusion models, leading to advancements in generating and manipulating image content [5, 11, 23]. However, applying these advancements to video editing has been relatively limited, stemming from inconsistencies inherent in the outcomes generated by text-to-image models. Within the domain of video editing and inpainting, a key objective is to ensure visual and temporal coherence in all frames of the edited/generated video. To this end, considering different applications such as video synthesis and video editing, different methods are proposed in the literature that can be categorized into three groups based on what they condition on: 1) Mask-guided methods [6, 21, 22, 25] in which deep neural models are conditioned on the binary mask of the input video. These methods are also called video inpainting methods in which the goal is to fill the masked (missing) area in the input video. Most of these methods are utilized for objects, watermarks, and logos removal, and video completion tasks. 2) Text-guided methods [3, 7, 18]: These approaches usually benefit from a diffusion model [11] conditioned by a text prompt. More specifically, this type of method requires detailed textual descriptions of both the original and target videos and then reconstructs the videos based on these descriptions for video synthesizing and editing purposes. 3) Multi-modal-guided methods: Recent approaches [1, 15, 24] introduce new stronger conditioning for full video generation, such as a style image, pose, depth, and sketch which leads to having more flexibility and control ability for video generation. However, they are not able to support localized editing of videos.

Despite the progress made in this area, existing methods still suffer from inconsistencies across frames of the generated/synthesized videos. Moreover, all the mask-guided methods assume that the mask by which their model is conditioned has a fixed shape and position in all the frames which is not generally the case in real-world video inpainting applications. Hence, this type of mask-guided method fails to handle mask images whose positions and geometric shapes vary across the frames. Last but not least, among diffusion-based consistent video editing techniques, almost all of them require either fine-tuning or training on large text-image or text-video datasets which practically is costly, time-consuming, and in some situations infeasible. To overcome these issues, inspired by the idea of extended attention layers [2, 3, 17], we design and develop a mask and text video editing framework by leveraging a pre-trained inpainting image diffusion model [11], allowing novel applications in mask guided video editing. We systematically redesign the forward path of the pre-trained diffusion model by replacing the self-attention modules with an extended version of attention modules without any additional training or fine-tuning. In this way, we ensure that the edited information will be consistent across all the video frames no matter what the shape and position of the masked area is. This approach allows us to use our framework for multiple video editing tasks such as object retargeting, object replacement, and object removal tasks while previous mask-

*This research was partially funded by NSERC Discovery Grant RGPIN-2021-04104 and Mila Tech Transfer, correspondence to: amirhossein.zamani@concordia.ca

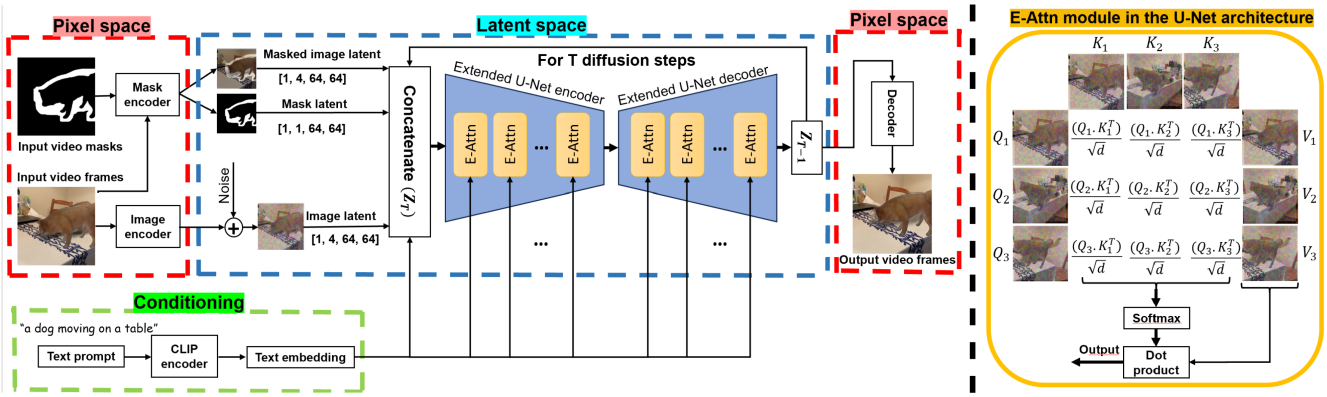


Figure 1. An overview of the diffusion process for temporal consistent video editing. To make the video editing process temporally consistent, we extend the U-Net architecture by replacing original attention modules used in [11] with Extended Attention modules

guided methods only focus on the object removal task.

## 2. Methodology

Stable Diffusion (SD) [11] is a well-known text-to-image diffusion model that adds/removes the noise to/from the latent (feature) space of the image, not the image itself during the diffusion process. This model can be conditioned on different modalities of signals such as mask, depth, pose, etc. By leveraging the mask-driven version of the text-to-image SD, which is called the inpainting stable diffusion model, we aim to design a video editing framework given a text prompt and a sequence of mask images corresponding to each frame of the input video. We emphasize that, unlike previous mask-guided inpainting methods [6, 21, 22, 25], our approach can deal with masks with arbitrary shapes, positions, and placements in each frame of the input video without any problem. Following the mask-generation strategy presented in [14], the inpainting diffusion model is resumed from an SD model and fin-tuned for another 200k steps with an additional conditioning signal that provides a mask corresponding to the input image. More specifically, compared to text-to-image stable diffusion models [10–12] which take as input the number of timesteps (diffusion steps) and the text embedding (obtained from a pre-trained text encoder [9]), the existing UNet model in the inpainting stable diffusion model has five additional input channels containing four for the encoded masked image and one for the mask itself. Specifically, these five channels are concatenated with the four ones (timestep and text embedding) used in the text-to-image stable diffusion model. Our approach is to first obtain an embedding representation of each frame and its corresponding mask in the input video by using a pre-trained variational autoencoder (VAE). Having the input frame and its mask, we then generate the masked image in which the area that we would like to perform the inpainting task on, has been masked out by a black color. Lastly, by concatenating three signals, we obtain the proper input for the inpainting SD model: 1) Noisy input frame latent: The added noise comes from the last diffusion step of the DDIM inversion stage [13] which is added to the embedding of the frame in the latent space, 2) Masked image latent which is obtained from the encoder of a pre-trained VAE model, and 3) Mask latent which is the down-sampled version of the VAE representation of the mask by using the nearest interpolation of the VAE representation (we suggest the reader to see Fig. 6 in the supplementary material that visually demonstrates how we achieve a proper input for the inpainting model). After feeding the proper input to the model, in the denoising diffusion stage, the UNet estimates the amount of existing noise in the input. Then, in each denoising step, the model tries to remove the estimated noise from the image latent and obtain a less noisy version of it while editing the masked area in the image by attending to the context of condition signals (mask and text) using existing attention layers in the UNet architecture.

The inpainting SD model gives us more flexibility to control the video content editing by leveraging a user-defined mask (determining a specific region in the input video for the inpainting task) and a text description. Moreover, it allows us to perform different video editing tasks presented in the next section by tuning the guidance scale (GS) parameter [4]. GS guides the model toward any of the two control signals we have: mask and text. However, applying the inpainting SD in a naive frame-by-frame manner leads to inconsistencies in frames of the edited frame. Instead, inspired by [2, 3, 17], we systematically redesign the forward path of the pre-trained inpainting diffusion model by replacing the self-attention modules with an extended version of attention modules that induces dependencies between frames. Note that we do not change the architecture of the existing U-Net in the SD model. We manipulate only the computation in the forward path of the self-attention layers. This happens by using several frames instead of one in the computation of self-attention modules to extract similar information or features. That is why our approach does

not add any additional training or fine-tuning. Then, having these extracted similar features, we enforce the model to edit (reconstruct) the video in a way that the regions in the frames that have similar features will be kept unchanged while the other parts of the frames will be changed according to the control commands (mask and text prompts). In this way, we ensure that the edited information will be consistent across all the video frames no matter what the shape and position of the masked area is. Fig. 1 (right) demonstrates a visual representation of how the forward path of the extended attention works in our framework. Fig. 1 (left) shows the whole diffusion process of our temporal consistent video editing technique. More specifically, for each diffusion step, similar to [3], we randomly select several frames and their corresponding mask images. Then, these pairs of masks and images are fed into a pre-processor algorithm explained above. Then, the extended attention layers in the U-Net architecture, showing in Fig. 1, extract similar features from the selected frames. This process is repeated for $T = 50$ diffusion steps.

## 3. Experiments and Results

To show the effectiveness of our temporal consistent approach, we perform three experiments: consistent retargeting, object replacement, and object removal. We present the comparative results of the proposed method in three subsections in the sequel. More specifically, we qualitatively compare the results of the proposed models, against ProPainter [25] and E2FGVI [6] (in the appendix) for object retargeting and object removal tasks. Also, we show the qualitative results for the object replacement task for different examples. All the experiments are performed on different videos from BANMo [20], DAVIS [8], and YouTube-VOS [19]. More results on these three tasks are presented in the Appendix.

**Video Object Replacement Task**. The goal of this task is to generate an edited video by replacing the foreground with another foreground that adheres to both text and mask prompts while keeping the background the same as the one in the original (input) video. Fig. 2 showcases the qualitative result of our method on video examples containing different animals and humans with different textual and mask control prompts. As shown, the model can faithfully replace the target object, determined by the mask prompt, with another object which is determined by the text prompt. To the best of our knowledge, our approach is the only one that is able to achieve this task with a mask-based guidance.

**Video Object Removal Task**. The goal of this task is simply to generate an edited video by replacing the foreground with the background. Almost all mask-guided methods in the literature only try to solve this task. Fig. 3 presents a qualitative comparison of our method against the ProPainter [25]. For this task, although our method obtains results with the same level of high fidelity compared to state-of-the-art, it still performs acceptably and much better in many other

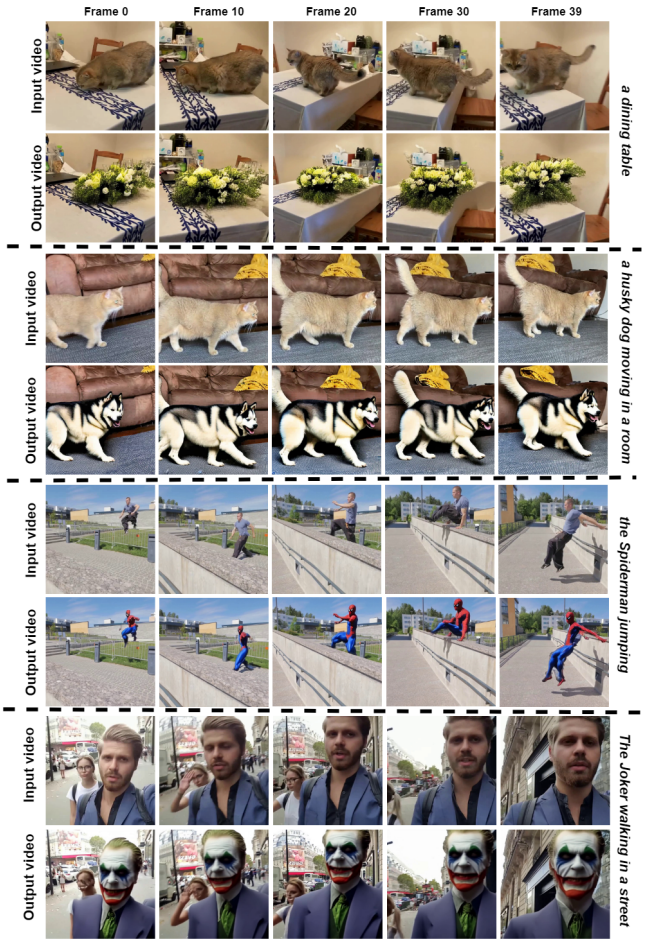


Figure 2. A qualitative results of the video object replacement task for different examples

methods in the literature, considering we only fine-tune the parameter guidance scale in our framework without changing any other components or parameters or any fine-tuning or training the model. This enlightens a way to propose a general framework that potentially not only could solve all the discussed tasks together in a unified pipeline, but also obtain state-of-the-art results in this area.

**Consistent Video Object Retargeting**. Object retargeting is a common task in graphics applications [20], that not only removes an object in the scene in a video or an animation sequence, but it replaces it with another object that performs the same action. This comes with significant challenges as now the action needs to be extracted from the data and applied to the new object. However, existing methods extract an object from a video and replace it with an object whose action may no longer be consistent with the video. For example: given a video of two objects (e.g. a dog and a cat) that have been overlaid on top of each other (see the second row of Fig. 4) and a sequence of mask images corresponding to all video frames, the goal of the consistent video object retargeting task is to keep the object which is on top, the

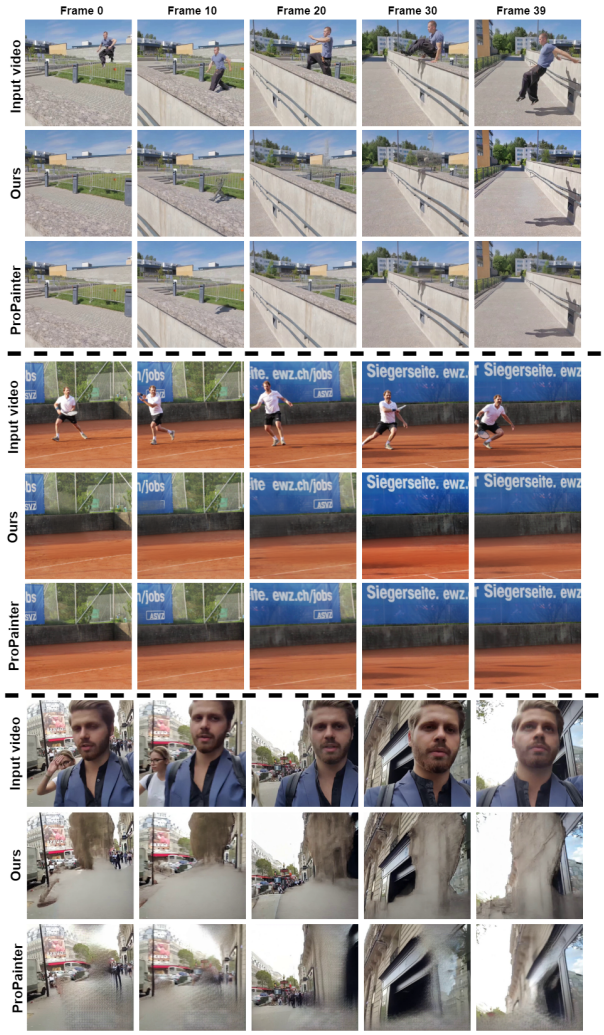


Figure 3. A qualitative comparison between ProPainter [25], E2FGVI [6], and our proposed method for the object removal task

*front object*, in the scene and instead remove the other object which lies behind the front object, *behind object* from the scene. More specifically, by processing the masked area, the model tries to remove the *behind object* from the scene in a way that everything in the background becomes temporal consistent across all frames. Our retargetted objects are created using [20]. Fig. 4 present a comparative analysis of the proposed against state-of-the-art methods including ProPainter [25] for a video of a dog and cat, focusing on qualitative video editing. In each column, which showcases the information belonging to a specific time in the input video, the first and second rows are the input masks and input frames for that specific time, and other rows demonstrate the result of the edited video when applying our method and ProPainter [25], respectively. The results show the superiority of our method to ProPainter [25] as there are some blurry regions in the results of Propainter [25] while this is not the case for ours (see more results and


Figure 4. A qualitative comparison between ProPainter [25], E2FGVI [6], and our proposed method for the object retargeting task

also comparison with E2FGVI [6] in the appendix). Similar to other mask-guided methods in the literature, the reason ProPainter [25] and E2FGVI [6] fail in this task is that all of them only can handle masks that have a fixed shape, position, and orientation across the frames. However, our method regardless of what the mask is tries to perform the inpainting tasks while achieving the temporal consistency at the same time.

## 4. Conclusions

This study introduces a temporal consistent method for video editing with mask and text guidance, relying only on a pre-trained in-painting diffusion model. The proposed approach allows one to use our framework without any additional training or fine-tuning to obtain competitive results on a common object removal task as well as a new object replacement task that prior approaches where existing methods are not viable. As for future work, we believe that the proposed approach can be used for a general set of tasks by enhancing its performance on the object removal task.

## References


[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 1

[2] 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. 1, 2

[3] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1, 2, 3, 4

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[6] 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. 1, 2, 3, 4, 7, 8

[7] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 1

[8] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3

[9] 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 *Proceesing of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[11] 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. 1, 2, 3

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

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

[14] 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

[15] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[17] 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. 1, 2

[18] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023. 1

[19] 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. 3

[20] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 3, 4

[21] Yongsheng Yu, Heng Fan, and Libo Zhang. Deficiency-aware masked transformer for video inpainting. *arXiv preprint arXiv:2307.08629*, 2023. 1, 2

[22] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*, pages 74–90. Springer, 2022. 1, 2

[23] 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. 1

[24] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 1

[25] 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. 1, 2, 3, 4, 7, 8

# Temporally Consistent Object Editing in Videos using Extended Attention

## Supplementary Material

## 5. Quanitative Evaluation

To numerically evaluate our method, we follow the criteria used by several state-of-the-art approaches in the literature [6, 25]: Structural similarity (SSIM) [16] index and peak signal-to-noise ratio (PSNR). These are methods for measuring the similarity between two images. The SSIM index can be viewed as a quality measure of one of the compared images, provided the other image is of perfect quality. We use both PSNR and SSIM metrics to evaluate the visual and perceptual similarity between input and output videos. More specifically, comparing our method with ProPainter [25] and E2FGvI [6] using the metrics mentioned above provides an insight into how similar the corresponding parts are in the input and output (edited) video frames. This is useful when we compare all the methods on the video object removal and video object retargeting task (explained in Section 3 of the main paper). Since in these two tasks it is desired to remove the masked-out areas and replace them with contents around the masked-out area in the input frames. As a result, it is important to know how effective the methods are in editing the masked-out object while keeping other areas in the input video unchanged. Note that we do not compare the results of our method with Propainter and E2FGvI for the video object replacement task as those methods cannot perform those tasks. We achieve better results with all metrics in the video object retargeting task, as indicated in Tab. 1. The superiority of the proposed method can also be visually verified by the qualitative results in Fig. 4 of the main paper. Regarding the results of the video object removal task, our method achieves results more or less the same level of fidelity compared to the state-of-the-art.

| **Methods** | **Removal Task** | | **Retargeting Task** | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| **E2FGVI** [6] | 18.5454 | **0.7510** | 19.7115 | 0.7782 |
| **ProPainter** [25] | 18.1411 | 0.7416 | 20.1209 | 0.7864 |
| **Ours** | **18.5871** | 0.7467 | **22.8837** | **0.8112** |

Table 1. A quantitative comparison between our proposed method, ProPainter [25], and E2FGVI [6]

## 6. Ablation Studies

To show the effectiveness of our design choices, we conduct ablation studies to analyze the underlying mechanisms and components of the proposed methodology by systematically evaluating the effects of removing or modifying specific elements. Fig. 5 demonstrates ablation experiments with different configurations. Note that we present the results of different ablation experiments on different video examples, and text-mask pair prompts to show the effectiveness and generalizability of our method. We present the ablation studies in four parts in the sequel.

**Different inpainting models**. As mentioned in the main paper, we leverage a pre-trained text- and mask-to-image inpainting diffusion model in our methodology. Among all these pre-trained inpainting stable diffusion models, we compare two of the most well-known models developed by StabilityAI [1] and RunwayML [2]. In this experiment, we qualitatively evaluate the performance of these two models in different video environments. In all examples, as shown in Fig. 5, the StabilityAI model generates more realistic images with higher quality. Rows #1-3 in the figure depict several frames of two edited videos resulting from the StabilityAI and RunwayML inpainting diffusion models on the same example with the same mask and text prompts. As shown, the edited video (the flower) is much more realistic than the video produced by the RunwayML model. Note that the depicted frames in Fig. 5 represent only one example to visualize the quality of these two models in our ablation studies. It is to be noted that our selection of models and examples is based on experimenting with different inpainting models with a sufficient number of examples to compare the quality and sanity of the edited images according to specific text prompts.

**Guidance scale**. We ablate the effect of the *guidance scale* (GS) parameter on the video editing qualitative results. Rows #5-7 of Fig. 5 show the effect of different GS's on the visual quality of the edited video. According to these results, on the one hand, when the guidance scale is set to a high value, the model strongly adheres to the text prompt. This often leads to images highly aligned with the provided text and mask image prompts, increasing fidelity to the input text (see row #5 of the figure). On the other hand, a lower guidance scale means the model gives more weight to the noise component during the diffusion process. This results in more diverse and creative outputs, but they are less faithful to the text prompt (see rows #6, 7 of the figure).

**Extended attention and historical frame information**. To demonstrate the effect of the extended attention module, we qualitatively compare two scenarios for several video examples.

- We replace the self-attention layers in our U-Net architecture with the extended version of the attention layer and use historical frame information to enforce the tem-

[1] Inpainting Stable Diffusion v2 by StabilityAI

[2] Inpainting Stable Diffusion v1.2 by RunwayML

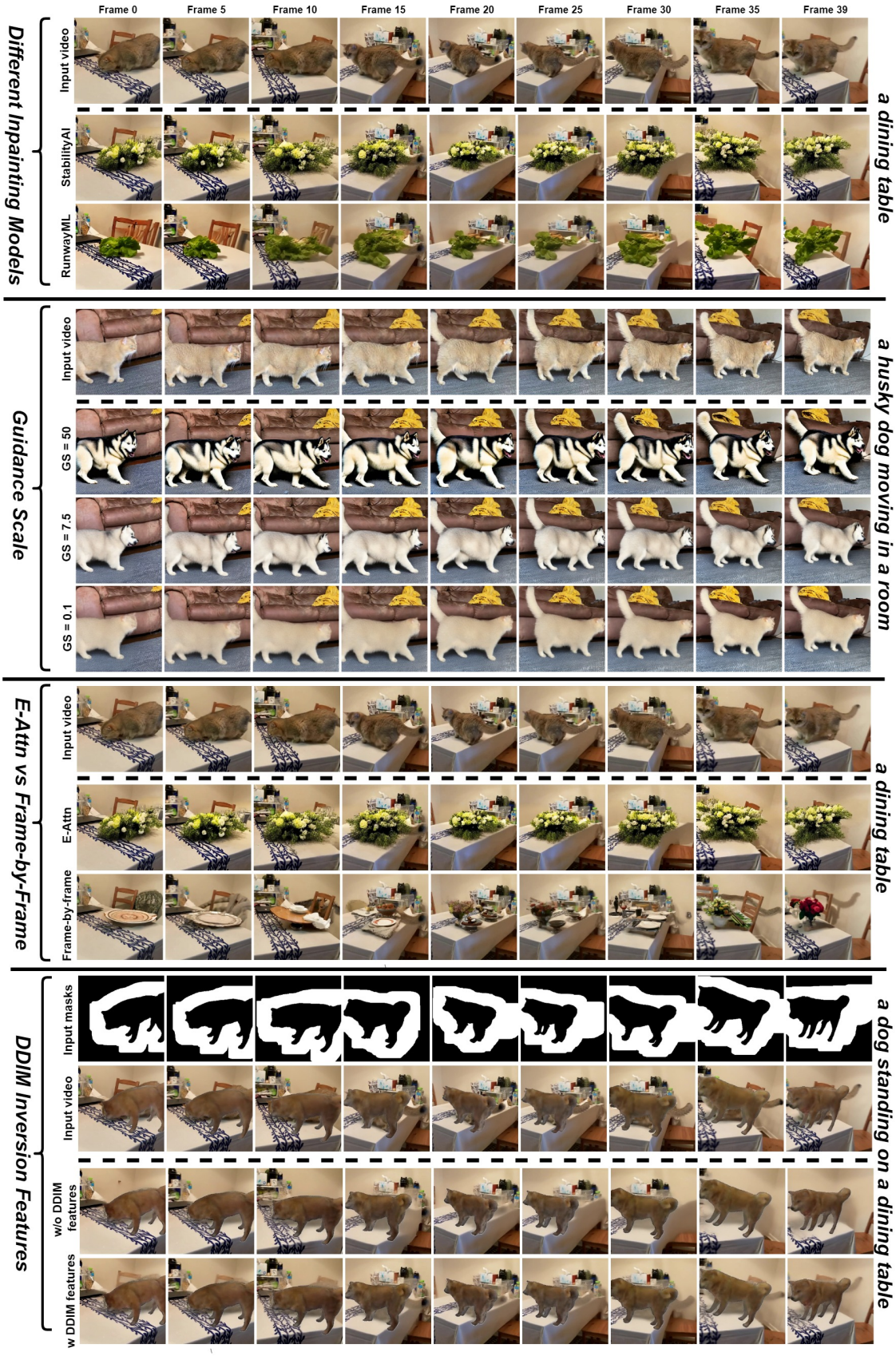


Figure 5. Ablation studies on different components and mechanisms of the proposed method.

poral consistency in the U-Net architecture (we refer to this scenario as *E-Attn* in Fig. 5).

- We apply the inpainting stable diffusion model in a naive frame-by-frame manner on the input video frames (we call this scenario *frame-by-frame* in Fig. 5). As shown in rows #9 and 10 of the figure, using the extended attention module results in visual consistency in the frames of the edited video while the frames are completely different and inconsistent in the frame-by-frame scenario.

**Features extracted from the DDIM inversion step**. We ablate the effect of incorporating the features extracted from the DDIM inversion step in our temporally consistent video editing procedure. The goal of this step is to obtain a noisy version of each frame by adding a small amount of noise to the less noisy version of the image (the pure image). Existing text-to-video editing methods such as TokenFlow [3] utilize these noisy versions of the frames for all diffusion steps obtained from the DDIM inversion process as knowledge to inform the model during the video editing step to keep the content of the edited video as close as possible to the input video. However, for the video inpainting tasks where both mask and text prompts contribute to the edited video, we empirically show that there is no need to incorporate all these noisy versions of the video frames. Instead, the only information required is the noisiest version of video frames (obtained from the last step of the DDIM inversion process) which will serve as an initial value for the denoising step during the editing process of our method. To perform this ablation study, we consider two scenarios.

- We remove DDIM inversion features and only consider text-guided, mask-guided, and unguided features during the inference stage.
- We concatenate DDIM inversion features with both guided and unguided features.

The last two rows of Fig. 5 depict the results of the video object retargeting task in these two scenarios. The results show that removing the DDIM inversion features improves the visual quality of the edited video significantly, while using the DDIM inversion features provides information about the frames of the input video which enforces the method to keep that information in the edited video as well. As a result, some objects/contents (such as the tail of the cat) from the input video can still be observed in the edited frames (the last row of Fig. 5) However, this is not a desired effect in applications such as video object removal or retargeting, where it is desired to remove the masked-out area and replace it with other objects in a way that the edited contents are visually consistent and coherent with the non-masked-out area.

## 7. Modifications to SD Model

In this section, we outline modifications we made to the text-to-image SD model [11] for temporally consistent object editing in videos.

**Using the inpainting SD model.** The existing text-to-image diffusion models such as SD [11] can only process a textual description as an external command given by a user to manipulate, edit, and synthesize an image. Therefore, to add more control ability to video editing applications, we replace the stable diffusion model recently leveraged by methods such as TokenFlow [3] with an inpainting SD model (described in Section 2 of the main paper) which can process both text and mask control commands. Note that this is carried out in the consistent video editing step while we use text-to-image in the DDIM inversion step. The proposed modifications to these two steps are described below.

**Leveraging only the last step of the DDIM inversion process.** While all the noisy frames obtained from the DDIM inversion step (explained in the previous section) are required to generate high-quality videos in recent video editing methods such as TokenFlow [3], we empirically found out that, for video inpainting tasks where both mask and text prompts contribute to the edited video, we need to use the noisy frames generated from only the last step of the DDIM inversion, not all steps (ablation results on this are presented in Fig. 5). Intuitively, this is due to the fact that using the mask control command limits the model to edit only the masked-out area in the input video. In this way, the extended attention layers ensure the consistency and coherence of non-masked-out areas by comparing corresponding contents in several random frames of the input video. However, existing methods such as TokenFlow [3], in addition to their consistency methods (no matter what they are), need to incorporate the noisy latent of each input frame to ensure temporal consistency and to keep the information in the non-masked-out areas unchanged as there is no control command other than a textual description. This modification drastically reduces the memory required for our method, as there is no need to compute and process noisy versions of all input video frames in all diffusion steps.

**Pre-processing input video frames along with text and mask prompts.** Since we replace the text-to-image SD model with the inpainting SD model, we need to provide a meaningful and reliable way to combine and feed the video frames, text, and mask prompts to the inpainting model. To this end, we follow the original implementation of the pre-processor algorithm [3] [4] for the inpainting SD model. We visually demonstrate how we achieve a proper input for the inpainting SD model in Fig. 6. A detailed explanation of all the pre-processing steps is presented in Section 2 of the main paper.

**Redesigning the forward path using the extended attention.** The core idea and main modification of our method is to systematically redesign the forward path of the pre-

[3] Pipeline stable diffusion inpaint
[4] Image processor

trained diffusion model by replacing the self-attention modules with an extended version of attention modules without any additional training or fine-tuning. Applying the extended attention operations on several frames of the input video ensures that the edited information will be consistent across all the video frames. Fig. 1 (right) of the main paper shows exactly how key, query, and value tensors in the extended attention operation are initialized using several randomly selected frames in the input video. This approach allows one to extract similar contents and features across multiple frames of the input video and to keep them unchanged when editing the masked-out area during the inference stage. Fig. 7 showcases the functionality of our approach at the inference stage. At a high-level overview, for each diffusion step, similar to [3], we randomly select several frames and their corresponding mask images. Then, these pairs of masks and images are fed into a pre-processor algorithm, demonstrated in Fig. 6, that processes and combines masks and images. Then, the extended attention layers in the U-Net architecture, shown in Fig. 1 of the main paper, extract similar features from the selected frames. This process is repeated for T = 50 diffusion steps to obtain an acceptable edited video.

## 8. Additional Results

We show expanded results from the main paper featuring more frames and an additional baseline (E2FGVI [6]) in Figs. 8 to 10.

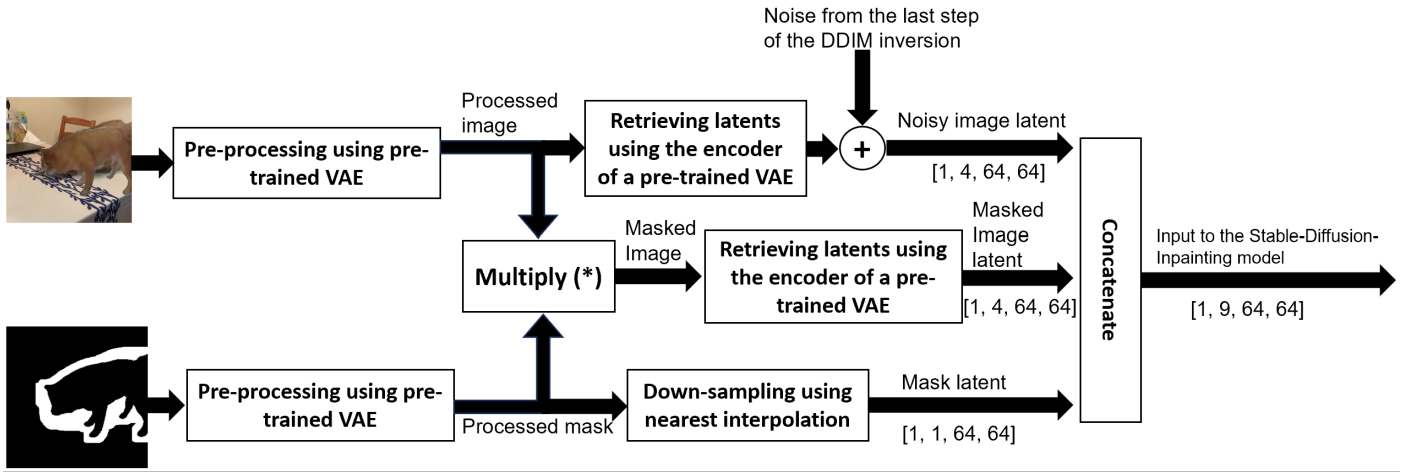


Figure 6. The mechanism of pre-processing the input video and its corresponding sequence of masks for the inpainting stable diffusion model

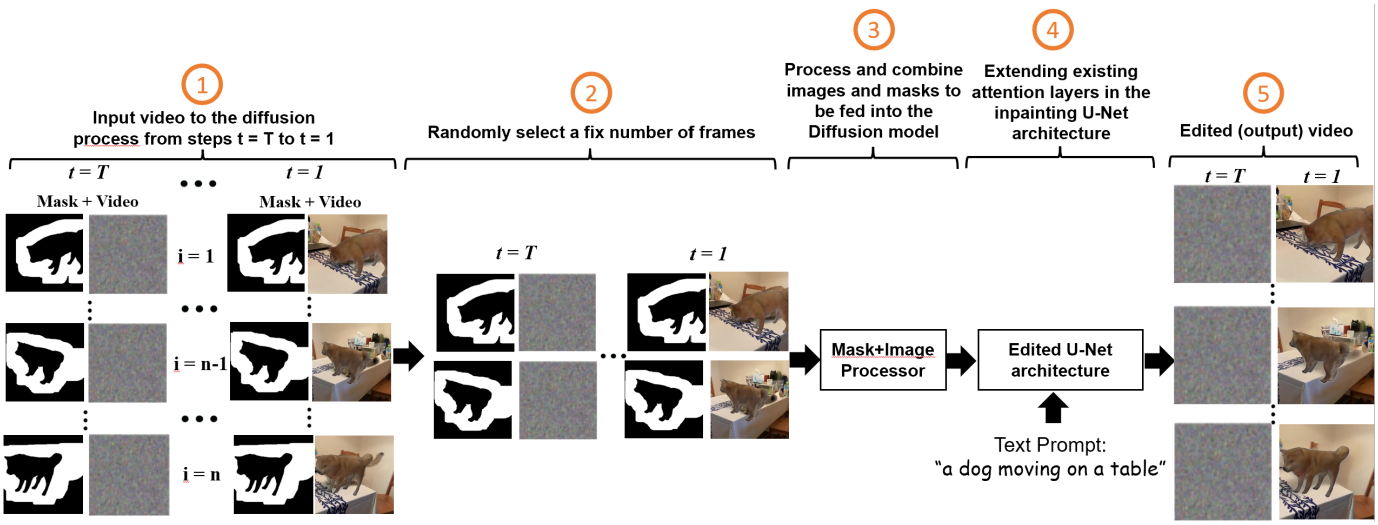


Figure 7. An overview of our pipeline at the inference stage

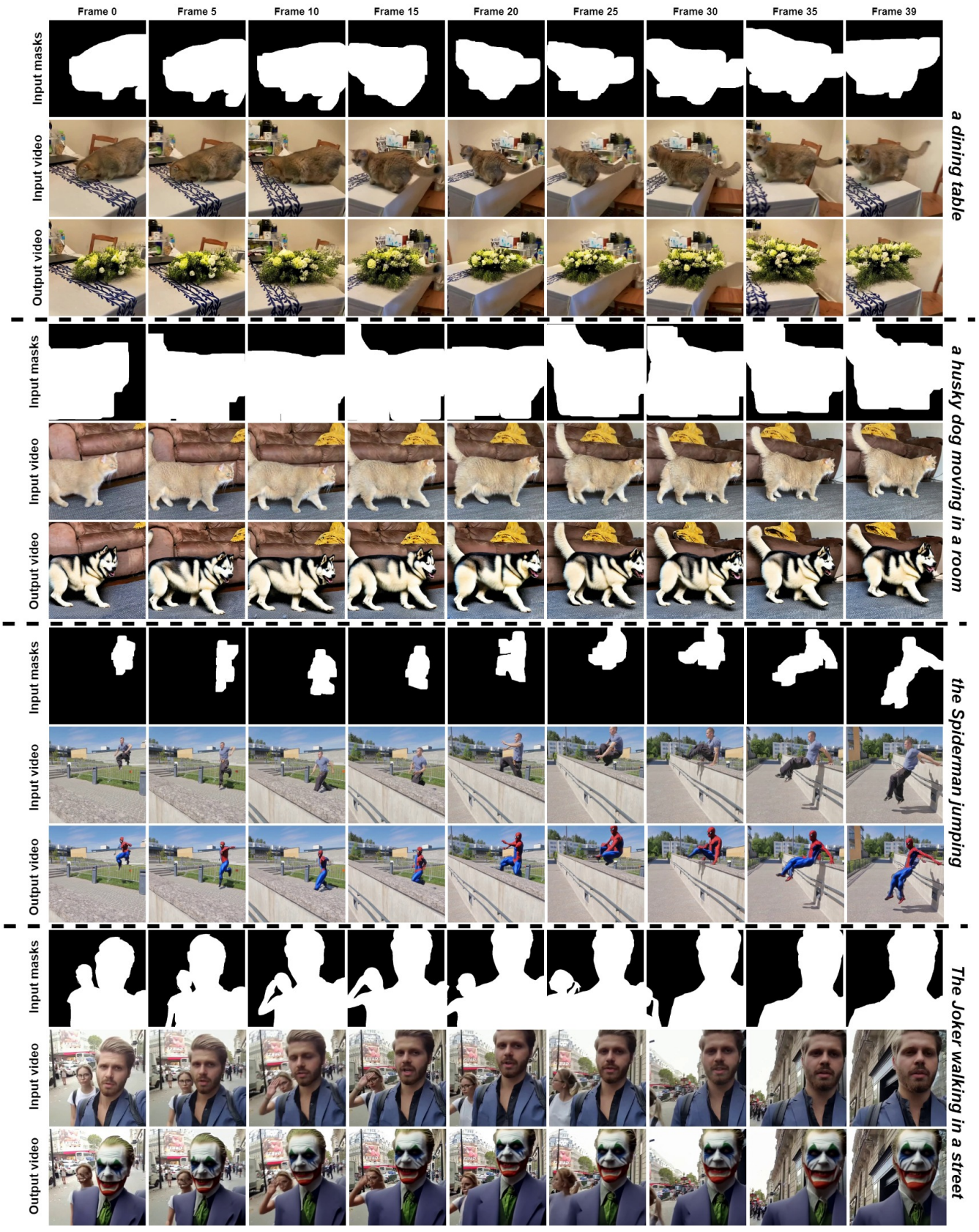


Figure 8. A qualitative results of the video object replacement task for different examples

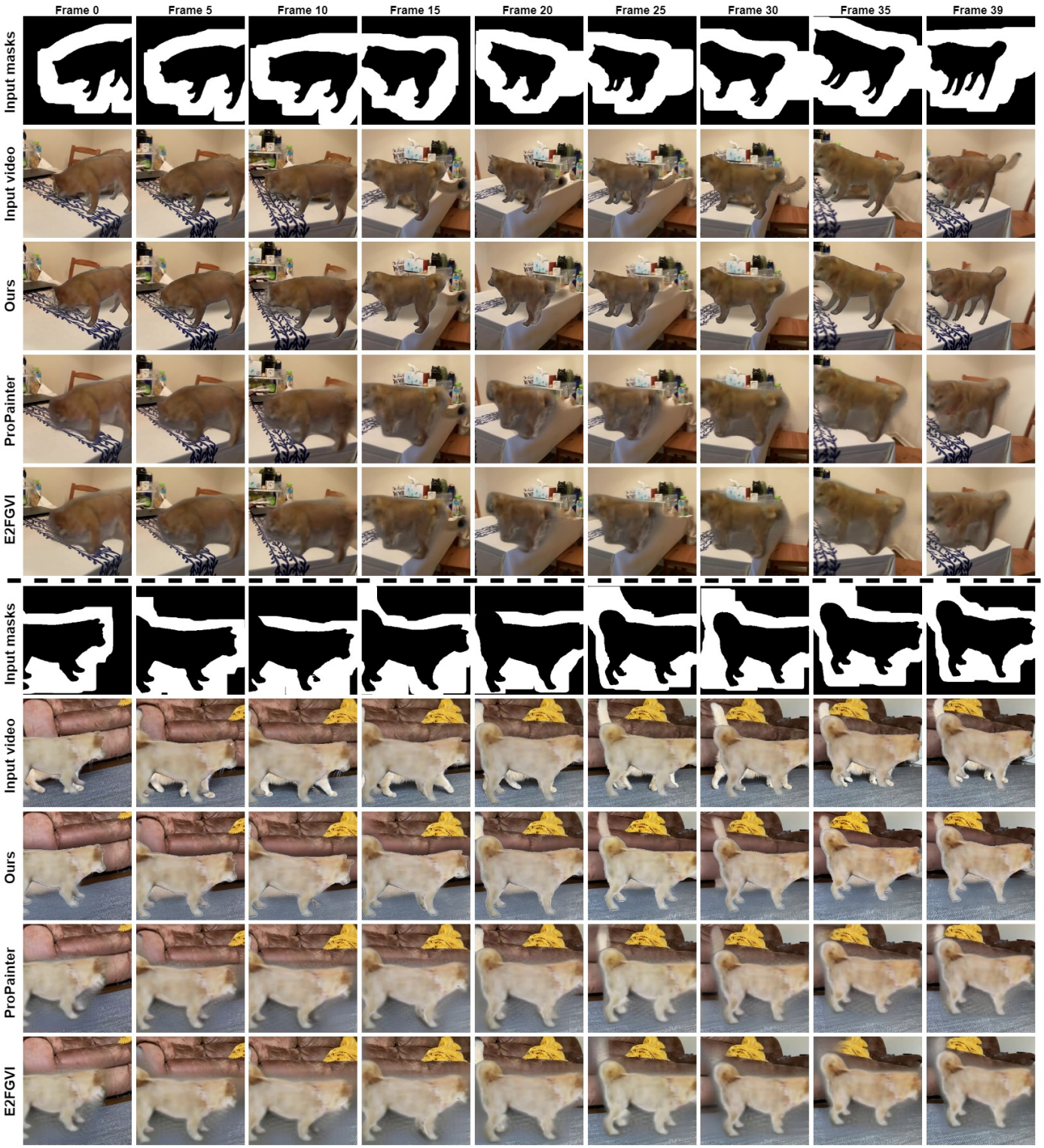


Figure 9. A qualitative comparison between ProPainter [25], E2FGVI [6], and our proposed method for the object retargeting task


Figure 10. A qualitative comparison between ProPainter [25], E2FGVI [6], and our proposed method for the object removal task