

1 A Dataset Details.

2 In the data preparation process, we synthesize video sequences within the Unreal Engine (UE)
3 environment. The virtual cameras are programmatically controlled to automatically focus on the
4 target object designated for removal. To simulate realistic motion, the cameras move randomly along
5 one of the three principal axes—X, Y, or Z—in the UE world coordinate system. The movement
6 distance along the chosen axis is adaptively adjusted according to the spatial scale and available
7 space within each specific scene, ensuring plausible camera trajectories without clipping or unnatural
8 transitions. By rendering the scene under these controlled camera motions, we obtain synchronized
9 video triplets consisting of the original unedited video, the corresponding masked video with the
10 target object occluded, and the edited video where the object is removed.

11 B Training Details.

12 During training, we use a batch size of 1 and randomly select a continuous sequence of 81 frames from
13 triplets of the original, masked, and edited videos as input. We extract and concatenate features from
14 Transformer layers block.5, block.15, and block.25(30 blocks in total), resulting in a feature tensor
15 of shape $[1, 28350, 4608]$, where 28350 corresponds to the flattened patch grid (e.g., $10 \times 15 \times 189$)
16 and 4608 is the combined feature dimension from the three layers. The Difference Mask Predictor
17 consists of a linear layer projecting the 4608-dimensional features to a 256-dimensional hidden
18 space, followed by a GELU activation and a final linear layer mapping to a single output channel.
19 The output, initially $[1, 28350, 1]$, is reshaped to $[1, 1, 10, 15, 189]$ and then upsampled via trilinear
20 interpolation to the full resolution $[1, 1, 81, 480, 720]$, matching the ground-truth difference mask size.
21 This design enables efficient patch-wise prediction of spatiotemporal difference masks, providing
22 fine-grained supervision for the video editing task.

23 C Potential impacts.

24 **Positive Impact.** This work is expected to significantly contribute to the field of video object
25 removal by providing a more robust and effective framework for erasing undesired regions in video
26 sequences. By greatly enhancing the quality, temporal consistency, and semantic coherence of the
27 inpainting results, it will not only push the boundaries of current video inpainting techniques but also
28 offer a solid foundation for future research and practical applications in video editing, surveillance
29 anonymization, and content restoration.

30 **Negative Impact.** Video inpainting technology, while powerful for restoring or editing visual
31 content, can also bring about negative impacts such as facilitating misinformation through realistic
32 content manipulation, undermining the credibility of video evidence in legal and forensic contexts,
33 and raising ethical concerns regarding privacy and consent. If misused, it may distort historical
34 records, violate intellectual property rights, or propagate biased or misleading visual narratives,
35 posing serious risks to information integrity and social trust.

36 D More Visualization Results.

37 In Fig. 1, we present more visualization results of our model.

38 E More Comparison Visualization Results.

39 In Fig. 2, we present more comparison results of our model with DiffuEraser [1], ProPainter [3], and
40 FuseFormer [2].

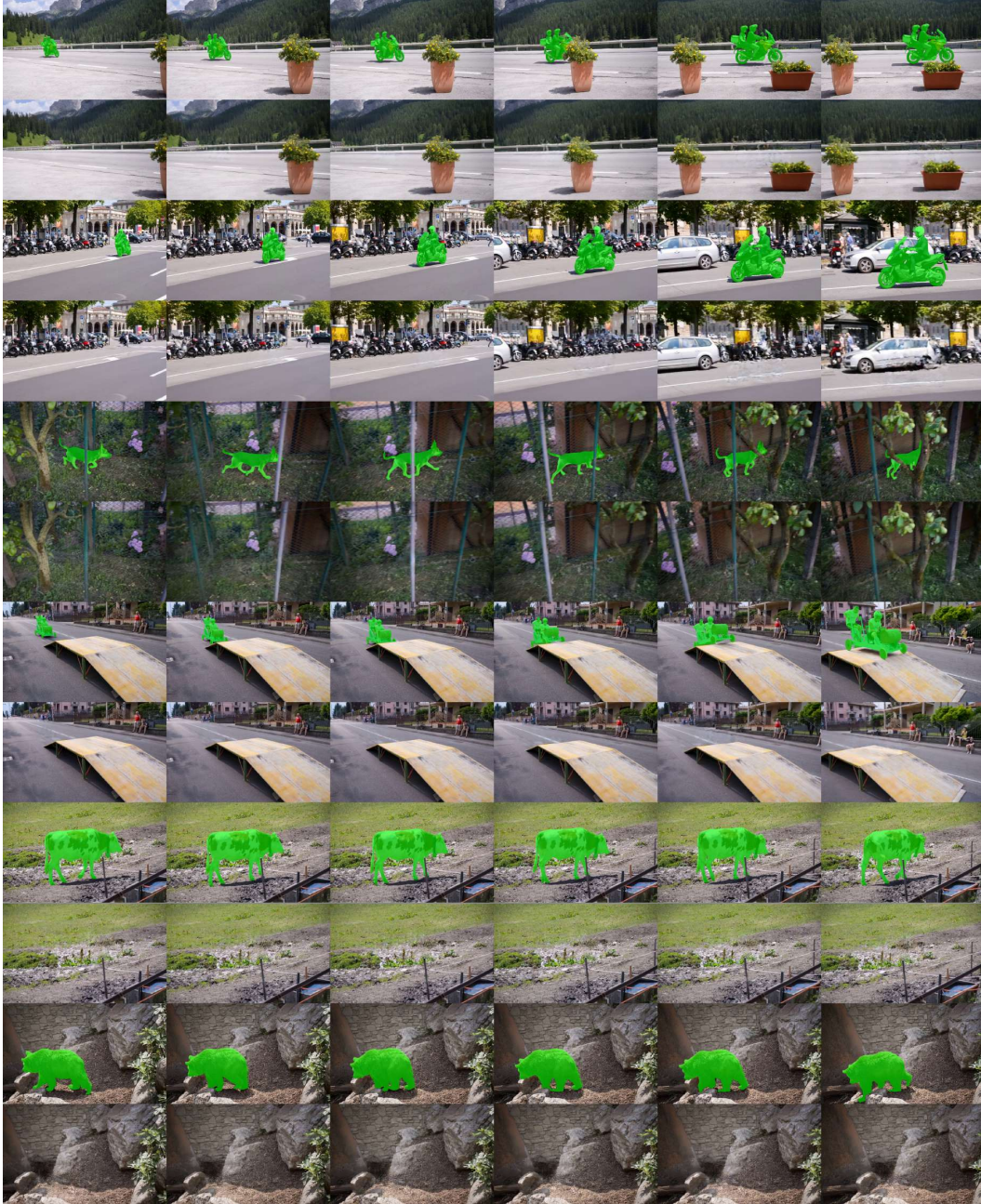


Figure 1: More visualizations.

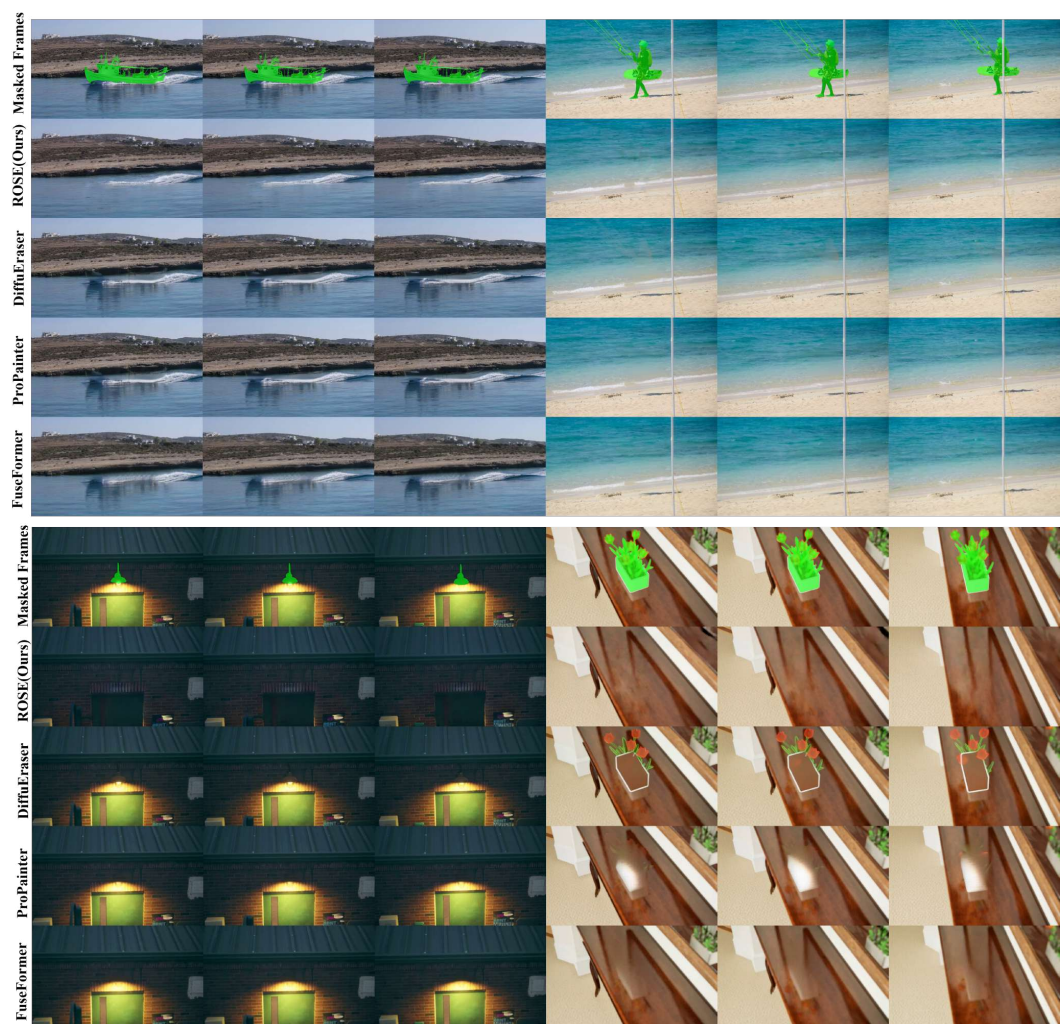


Figure 2: More comparisons.

41 **References**

- 42 [1] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffuener: A diffusion model for video inpainting.
43 2025.
- 44 [2] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai,
45 and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In
46 *ICCV*, 2021.
- 47 [3] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. Propainter: Improving propagation
48 and transformer for video inpainting. In *ICCV*, 2023.