

UniPaint: Unified Space-time Video Inpainting via Mixture-of-Experts

Supplementary Material



Figure A1: More results of our method. Additional qualitative results of our method applied to various inpainting scenarios, including object removal, environment swapping, outpainting, and temporal inpainting(interpolation). These examples demonstrate the flexibility of our method across diverse scenarios

Overview

This supplementary material provides additional details and insights to further elaborate on various aspects of the proposed method. The content is organized as follows:

- **Experiment Details:** Detailed information about the training and evaluation procedures can be found in Appendix A.
- **More Qualitative Results:** In Appendix B, we showcase an expanded set of qualitative experiments, highlighting the flexibility and consistency of our approach.
- **More Comparative Analysis:** Beyond the qualitative comparisons presented in the main paper, Appendix C includes further analyses focusing on marginal and temporal inpainting tasks.

A Experiment Details

Our model is trained using a two-stage procedure:

- (1) **Initial Training:** The model is first trained on the WebVid-10M dataset [1] for 5 epochs using a learning rate of 1×10^{-4} .
- (2) **Fine-Tuning:** Fine-tuning is conducted on the YouTubeVOS dataset [6] for 10 epochs with a reduced learning rate of 1×10^{-5} .

We utilize the AdamW optimizer [5] for both stages of training. The process is performed on 8 NVIDIA A100 GPUs over approximately 3 days. All ablation studies follow the same training configuration for consistency.

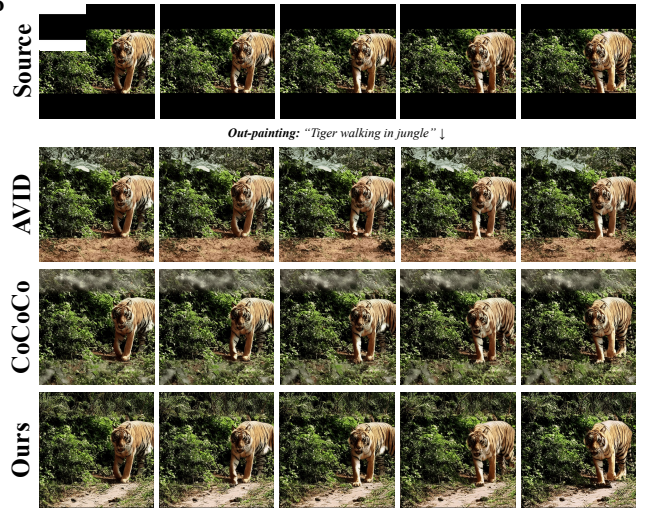


Figure A2: Comparative analysis of space-time video outpainting. We compare our method against AVID [9] and CoCoCo [10] for outpainting. Our method achieves the most realistic and coherent outpainting, with superior alignment, texture consistency, and scene continuity compared to AVID and CoCoCo.

For inference, UniPaint operates in float16 precision and requires 30 GB of GPU memory. Processing a single video clip takes 69 seconds on one NVIDIA A100 GPU.

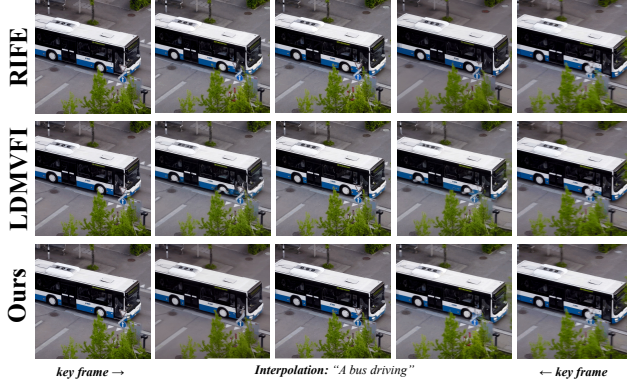


Figure A3: Comparative analysis on temporal inpainting. Key frames are provided at the beginning and end, with interpolated frames shown for RIFE [3], LDMVFI [2], and our method. RIFE shows blurriness and inconsistent details, while LDMVFI exhibits better temporal coherence but introduces blending artifacts and lacks sharpness. *UniPaint* achieves the most realistic and consistent temporal inpainting, preserving fine details, sharp edges, and seamless transitions across all frames.

B Qualitative Results

As illustrated in Fig. A1, we present additional qualitative results demonstrating the capabilities of our method across diverse inpainting scenarios. *UniPaint* exhibits strong adaptability and maintains consistent performance across a variety of inpainting tasks.

Object Removal. Early works on video inpainting often focused on object removal as a primary task [4, 7, 8]. While modern diffusion models provide greater generative flexibility, our method retains the ability to perform effective object removal. By applying appropriate masks and providing textual prompts that describe the desired background, *UniPaint* can efficiently eliminate unwanted objects from video sequences while maintaining spatial and temporal consistency.

Environment Swap. Environment swapping can be considered a specialized case of outpainting. By selecting the complement of the target region as the editing area, our method enables seamless integration of a foreground object into a custom background. Using prompts that describe the new environment, *UniPaint* accurately modifies the scene, ensuring that the object appears naturally within the specified setting.

C Qualitative Comparisons

We further conduct more comparative analysis against various inpainting models, as shown in Figs. A2 and A3.

Out-painting. For out-painting, we compare our method with AVID [9], CoCoCo [10]. As shown in Fig. A2, our method significantly outperforms both AVID and CoCoCo. AVID exhibits noticeable artifacts and blending issues in the outpainted regions, failing to maintain texture and scene consistency. CoCoCo produces more coherent outputs than AVID but lacks fine-grained alignment with the original scene, resulting in less natural extensions. In contrast,

our method generates sharp, realistic, and seamlessly integrated outpainting results, preserving both structural and textural fidelity to the original scene.

Temporal inpainting. For temporal inpainting, we compare our method with RIFE [3] and LDMVFI [2]. As shown in Fig. A3, our method achieves the best interpolation quality among the evaluated models. RIFE outputs suffer from blurriness and inconsistent details, particularly at object edges and in motion dynamics. LDMVFI demonstrates better temporal coherence than RIFE but introduces blending artifacts and lacks sharpness in reconstructed details, like the wheels of the bus. Our approach produces the most consistent and realistic temporal inpainting, maintaining fine details, sharp edges, and seamless transitions across frames, ensuring both visual fidelity and temporal smoothness.

References

- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1728–1738.
- [2] Duolikun Danier, Fan Zhang, and David Bull. 2024. LDMVFI: Video Frame Interpolation with Latent Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 2 (March 2024), 1472–1480. doi:10.1609/aaai.v38i2.27912
- [3] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. 2022. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*. Springer, 624–642.
- [4] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. 2019. Deep video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5792–5801.
- [5] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [6] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. 2018. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327* (2018).
- [7] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. 2019. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3723–3732.
- [8] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. 2020. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI*. Springer, 528–543.
- [9] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. 2024. AVID: Any-Length Video Inpainting with Diffusion Model. *arXiv:2312.03816 [cs.CV]* <https://arxiv.org/abs/2312.03816>
- [10] Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Kam-Fai Wong, and Lei Zhang. 2024. CoCoCo: Improving Text-Guided Video Inpainting for Better Consistency, Controllability and Compatibility. *arXiv:2403.12035 [cs.CV]* <https://arxiv.org/abs/2403.12035>