

ReCapture: Generative Video Camera Controls for User-Provided Videos using Masked Video Fine-Tuning

Supplementary Material

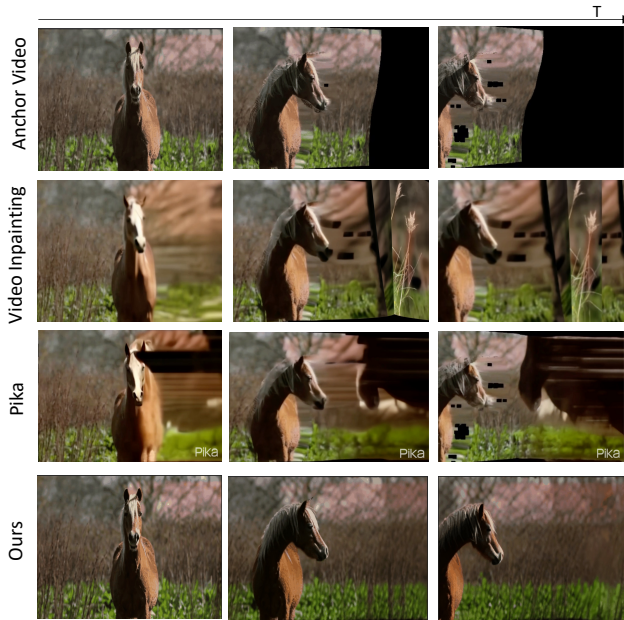


Figure 9. Comparisons with inpainting methods.

Our video results, including the gallery and ablation studies, can be found at cvpr25-submission-1616.github.io. Alternatively, you can open the ‘index.html’ file in your browser from the supplementary material folder path to access the website offline.

6. More Related Work

Personalization of Video Diffusion Models. At this stage the problem of personalization of image generative models has been well explored in the last several years, with work on subject-driven generation [4, 5, 11, 12, 18], style-driven generation [10, 14, 17], style+subject-driven generation [13] and image-level personalization for inpainting [15]. The research direction of personalization of video models [6] is more sparse, albeit with important recent work such as Dreamix [8] which proposed to finetune video models on a given video, Still-Moving [3] which mitigates the need for customized video data by elevating a customized image models to the video domain using spatial and temporal adapters and Movie Gen [9] which proposes directly training conditioning pathways for video models. Our method targets a wholly different application than this body of work, although these methods are important and related.

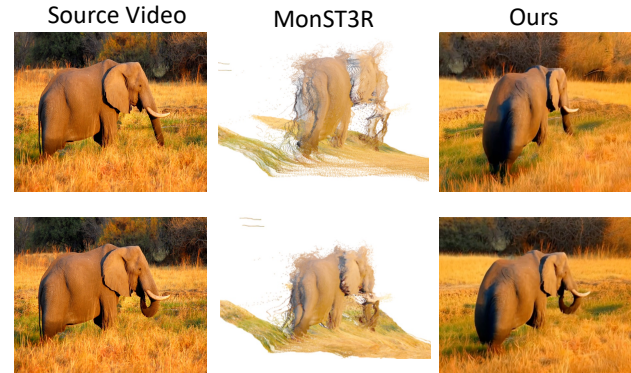


Figure 10. Comparisons with 4D reconstruction method MonST3R [19].

7. Comparisons with Video Inpainting

In the first stage, we obtain the anchor video along with a sequence of masks that represent the invalid regions of the video. A natural approach is to follow the methods used in Lumiere [1] and stable diffusion inpainting by leveraging an inpainting model, which uses the concatenation of the masks and the input video as its condition. We train such an video inpainting model using SVD [2] following Lumiere [1]. To evaluate the effectiveness of our approach, we compare our mask-video fine-tuning method against this standard video inpainting model and Pika [7] (an inpainting tool). As demonstrated in Fig 9, our method achieves significantly better visual quality compared to Pika and the standard video inpainting model. Additionally, our approach outperforms both Pika and the typical video inpainting model on the Vbench dataset as shown in Table 4.

8. Comparisons with 4D Reconstructions

We compare our method with 4D reconstruction approaches in Table 2 of the main paper and provide visualization comparisons with the latest state-of-the-art method, MonST3R [19], in Fig.10. As illustrated in Fig.10, MonST3R exhibits significant blurring due to the rendering trajectory being far from the training views. Similarly, most 4D reconstruction methods struggle to generalize beyond the field of view present in the original video. In contrast, our method leverages the strong prior of a video diffusion model, enabling it to produce high-quality novel views even with substantial camera movements.

Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
Video Inpainting	83.48%	86.25%	72.45%	81.06%	49.25%	37.72%	58.69%	79.68%
Pika	83.26%	84.96%	71.02%	80.45%	49.46%	39.08%	57.26%	79.68%
Ours	88.53%	92.02%	91.12%	98.24%	49.03%	57.35%	64.75%	82.07%

Table 4. Comparisons with video inpainting methods, including the video diffusion model and the Pika video inpainting tool.

Methods	Video Quality	Camera Following	Original Motion
Ours	96%	72%	92%
Genreative Camera Dolly [16]	4%	28%	8%

Table 5. User Studies for Visual Quality, Camera Following and Original Motion.

9. User Studies

We use 35 videos that were used for VenBench in the human evaluation. The survey is conducted on Amazon Mechanical Turk. To evaluate video quality and camera trajectory alignment(Camera Following), we present two videos from different methods in a random sequence and ask annotators to indicate which one has better quality and aligns more accurately with the provided camera trajectory. Additionally, to evaluate whether the subject and scene motion from the original video is preserved in the new video(Original Motion), we ask annotators to determine which video better maintains the original motion. As shown in Table 5, our method receives significantly higher human preference across both evaluation aspects.

References

- [1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2, 1
- [2] 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. 8, 1
- [3] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *arXiv preprint arXiv:2407.08674*, 2024. 4, 1
- [4] Wenhui Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ArXiv*, abs/2208.01618, 2022. 1
- [6] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1
- [7] Pika Labs. Pika labs, 2024. Accessed: 2024-11-20. 1
- [8] 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. 1
- [9] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1
- [10] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024. 1
- [11] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 4, 1
- [12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024. 1
- [13] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2025. 1
- [14] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: text-to-image generation in any style. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 66860–66889, 2023. 1
- [15] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. Realfill: Reference-driven generation for authentic image completion. *ACM Transactions on Graphics (TOG)*, 43(4):1–12, 2024. 1
- [16] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. *arXiv preprint arXiv:2405.14868*, 2024. 2, 6, 7

- 753 [17] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui
754 Qin, and Anthony Chen. Instantstyle: Free lunch towards
755 style-preserving in text-to-image generation. *arXiv preprint*
756 *arXiv:2404.02733*, 2024. 1
- 757 [18] Hu Ye, Jun Zhang, Sibor Liu, Xiao Han, and Wei Yang. Ip-
758 adapter: Text compatible image prompt adapter for text-to-
759 image diffusion models. *arXiv preprint arXiv:2308.06721*,
760 2023. 1
- 761 [19] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jam-
762 pani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-
763 Hsuan Yang. Monst3r: A simple approach for estimat-
764 ing geometry in the presence of motion. *arXiv preprint*
765 *arxiv:2410.03825*, 2024. 8, 1