

Reshoot-Anything: A Self-Supervised Model for In-the-Wild Video Reshooting

Supplementary Material

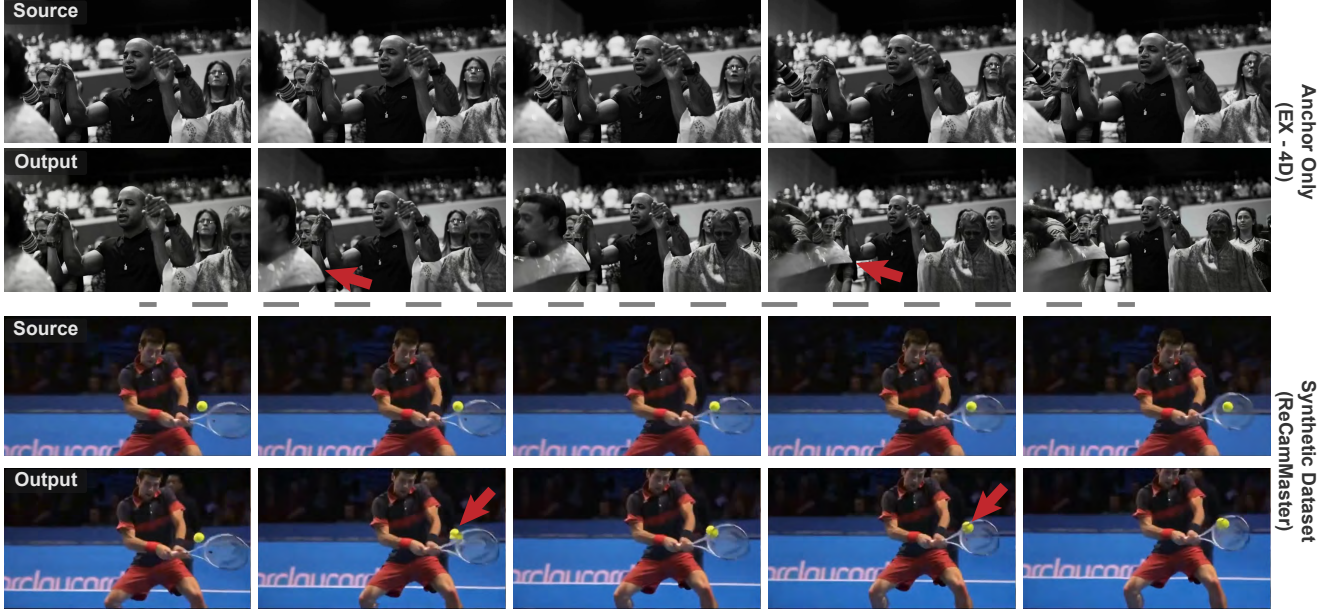


Figure 1. **Motivation: Illustrating Failure Modes in Video Reshooting.** This figure highlights common limitations of existing video reshooting methods, motivating our approach. **(Rows 1-2) Anchor-Only Artifacts:** Given a high-quality source video (Row 1), an anchor-only method like EX-4D [3] (shown in Row 2) often generates significant ghosting and content inconsistencies. This occurs when the model relies solely on an imperfect anchor video (V_a) for geometric guidance and struggles to hallucinate complex details not fully represented in V_a . **(Rows 3-4) Loss of Detail in Synthetic-Data Models:** For a given source video (Row 3), synthetic-data trained models, such as ReCamMaster [1] (Row 4), can fail to capture intricate spatio-temporal details and object dynamics. In this example, ReCamMaster incorrectly deforms the tennis ball as it moves across the racket. *Note on ReCamMaster:* While ReCamMaster demonstrates impressive capabilities in general camera-controlled video generation, this result (obtained from the demo videos on their official project website) highlights that challenges with intricate object dynamics persist even in their best demonstrations, and similar artifacts are evident in their other results.

In this supplementary material, we provide comprehensive implementation details, extended experimental results, and a discussion of potential applications for our video reshooting framework.

1. Motivation

Achieving photorealistic video reshooting requires simultaneously maintaining precise geometric alignment with a novel camera trajectory while preserving the intricate textures and dynamic content of the original source video. Existing approaches often struggle to balance these requirements, leading to characteristic failure modes illustrated in Figure 1. Methods that rely solely on sparse or imperfect anchor videos for conditioning, such as EX-4D [3], are forced to hallucinate missing texture information without ground truth, frequently resulting in deformed textures or severe ghosting artifacts. Conversely, models trained pri-

marily on synthetic datasets, like ReCamMaster [1], often exhibit a domain gap when applied to real-world footage, failing to capture complex physical dynamics and fine-grained details. These limitations motivate our approach, which utilizes a self-supervised framework designed to learn directly from real-world monocular videos and explicitly fuses high-fidelity source textures with anchor-based geometric guidance.

2. Implementation Details

2.1. Diffusion Transformer Configuration

Our diffusion transformer is built upon the **Wan2.2-12V 14B** model [10]. This base model employs a Mixture-of-Experts (MoE) design, featuring distinct parameter sets specialized for high-SNR (low-noise) and low-SNR (high-noise) regions of the diffusion trajectory. The architecture functions as a Latent Diffusion Model (LDM), per-

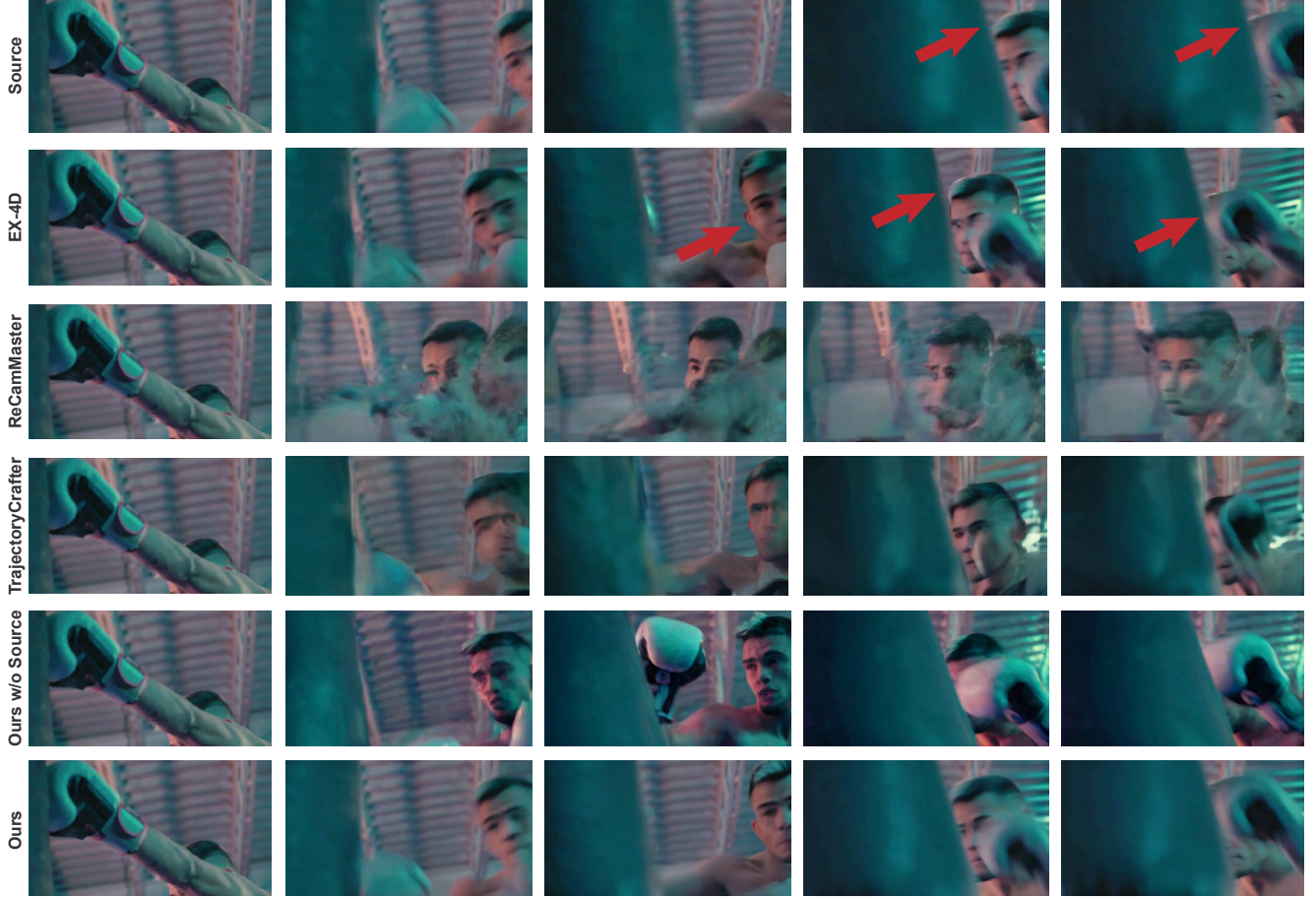


Figure 2. **Extended Qualitative Comparisons.** Each row displays sample frames from generated videos. Arrows indicate characteristic artifacts in baseline methods, such as loss of fine detail, blurring, or texture distortion. In contrast, our approach consistently demonstrates superior fidelity, accurately reproducing small details and effectively preserving intricate textures from the source video.

forming all video generation within a compressed latent space. Videos are encoded into a latent representation $z \in \mathbb{R}^{C_L \times T_L \times H_L \times W_L}$ by a causal 3D VAE with a U-Net backbone. This VAE temporally compresses the input, mapping the input video frames to T_L latent frames.

The native Wan2.2-I2V model is designed to condition on an image-reference latent, along with a binary mask indicating visible frame locations. Its standard conditioning pathway concatenates this image-reference latent, a C_M -channel binary mask, and a C_L -channel noisy latent along the channel dimension. This results in a tensor of shape $\mathbb{R}^{(2C_L+C_M) \times T_L \times H_L \times W_L}$, which is then patchified and processed by the DiT blocks utilizing 3D Rotary Positional Embeddings (3D-RoPE).

For our video reshooting task, we adapt this scheme to integrate both anchor and source video information:

Anchor Conditioning Stream The VAE-encoded anchor latent z_a is concatenated with a C_L -channel noise latent z_n and a C_M -channel downsampled binary mask M_a . This forms a tensor representing the anchor conditional input.

Source Conditioning Stream The VAE-encoded source latent z_s is duplicated along the channel dimension (replacing z_n) and then concatenated with an all-ones C_M -channel mask M_s .

These two conditional inputs are temporally concatenated, leading to a token sequence that is twice as long as standard inference (i.e., $2 \times T_L$ latent frames). We share the parameters of the patchify blocks across both pathways.

To distinguish the positional context of the appended source tokens, we apply a constant offset to their 3D-RoPE along the temporal dimension within the DiT blocks. This offset magnitude is set to 50, which significantly exceeds our maximum number of training latent frames ($T_L = 20$). This decoupling allows us to flexibly change the length of generated videos during inference. The self-attention layers within the DiT blocks attend jointly over all concatenated anchor and source tokens, enabling efficient cross-video knowledge transfer. Finally, we apply rank-512 LoRA fine-tuning to all attention and fully connected layers.

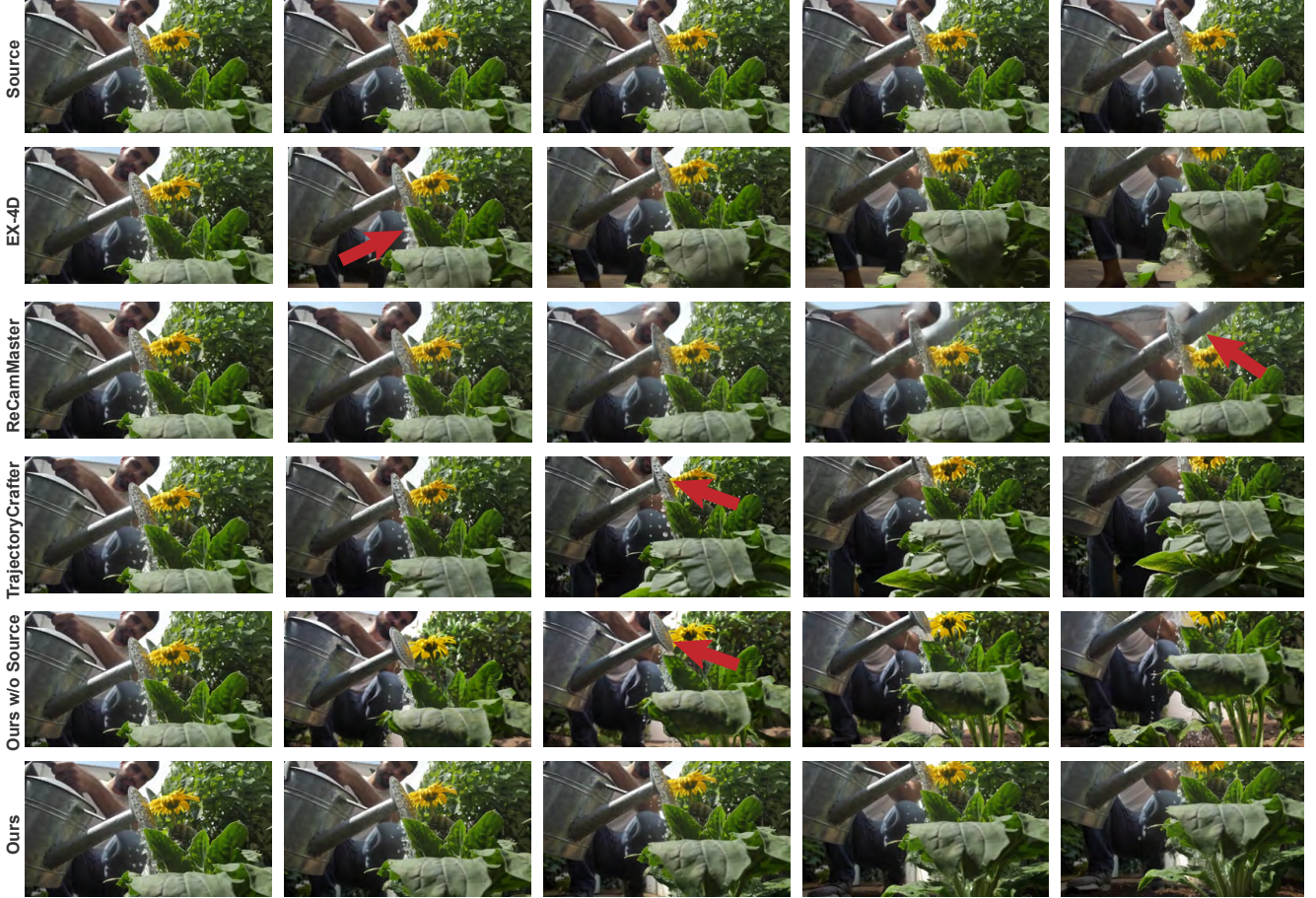


Figure 3. **Extended Qualitative Comparisons.** Each row displays sample frames from generated videos. Arrows indicate characteristic artifacts in baseline methods, as noted previously. Our approach maintains robust geometric fidelity and superior perceptual quality.

2.2. Augmentations and Ablation Setup

As introduced in the main paper, we implement several technical augmentations designed to enhance model robustness and visual quality.

Auxiliary Loss To ensure the source token pathway actively retains meaningful content, we apply an L1 reconstruction loss between the output tokens corresponding to the source video and the VAE-encoded clean source latent z_s . This loss is weighted by a factor of 0.1.

Fluorescent Background Anchor In standard anchor video generation, regions representing new viewpoints or disocclusions are filled with a black background. However, for dark scenes, the model struggles to distinguish masked regions from actual dark content. We found that replacing the standard black background with a high-contrast fluorescent pink color provides a distinct signal for the model, making the boundary of missing information unambiguous.

Random Query Our default anchor generation warps the first frame of the source video ($V_s[0]$) to create V_a . However, to make the training process more robust to diverse tracking conditions, we introduce an augmentation where

the reference frame for dense tracking and warping is randomly selected within the source video ($V_s[t]$). This prevents the model from developing a bias towards early frames and encourages sustained attention to geometric guidance throughout the sequence.

3D Noise in Anchor A potential issue arises when the synthesized anchor video (V_a) closely resembles the target video (V_t), tempting the model to directly copy texture from V_a instead of routing from V_s . To suppress V_a 's texture information while preserving its 3D geometric guidance, we inject Gaussian noise into the RGB values of the reference frame ($V_s[t]$) *before* it is forward-warped. The noise magnitude is sampled uniformly between $[0, 0.5]$ per channel. Unlike injecting noise *after* warping, pre-warping noise ensures that the noise itself moves coherently with the underlying 3D structure. This forces the model to rely on V_s for high-fidelity content while still inferring motion from V_a .

2.3. Anchor Generation for Inference

While our self-supervised training pipeline utilizes efficient 2D warping, during inference we utilize an explicit 3D projection approach to define the target camera trajectory. We



Figure 4. **Extended Qualitative Comparisons.** Additional examples demonstrating our model’s ability to consistently reproduce fine details and preserve intricate textures from the source video across diverse scenes.

extract dense geometric information from the source video using state-of-the-art monocular depth and camera estimation models [4, 5, 11]. Using these depth maps and camera parameters, we unproject the source video pixels into 3D world space, resulting in a consistent, colored point cloud. Finally, to generate an anchor frame corresponding to a target camera pose, we re-render this 3D point cloud from the novel viewpoint, utilizing the input cameras to cancel out the original camera motion.

2.4. Evaluation

Dataset. We constructed a final set of 100 videos, sub-sampled from 1,000 stratified examples from the Opensora-mixkit dataset [7]. Videos were randomly assigned 1 of 10 predefined camera motion trajectories. To ensure balanced movement representation, we calculated the mean value of the generated anchor mask (M_a) for each video, grouped them by trajectory, and selected the 10 videos nearest to their respective group median. This mitigates outliers with extreme mask presence while preserving the underlying semantic distribution.

Metrics. We evaluate along three key dimensions:

- *Camera Accuracy:* We measure Rotational Error (RotErr) and Translational Error (TransErr) following [2]. We extract camera poses using [5], align the first frame, and calculate relative poses between the generated and ground-truth trajectories. TransErr is the sum of squared L2 distances, and RotErr is the sum of angular differences.
- *View Synchronization:* To assess perceptual similarity, we calculate the Fréchet Video Distance (FVD-V) [9] between the source videos and generated outputs. For semantic synchronization, we use CLIP-V, the average frame-wise CLIP cosine similarity at matching timestamps. For fine-grained spatial alignment, we use Matching Pixels (Mat. Pix) via the GIM model [8], counting geometrically-verified (RANSAC) feature matches that exceed a confidence threshold of 0.5.
- *Video Quality:* We utilize the VBench benchmark [6] for perceptual quality. To evaluate temporal consistency, we compute CLIP-F, the average CLIP similarity score of adjacent generated frames.

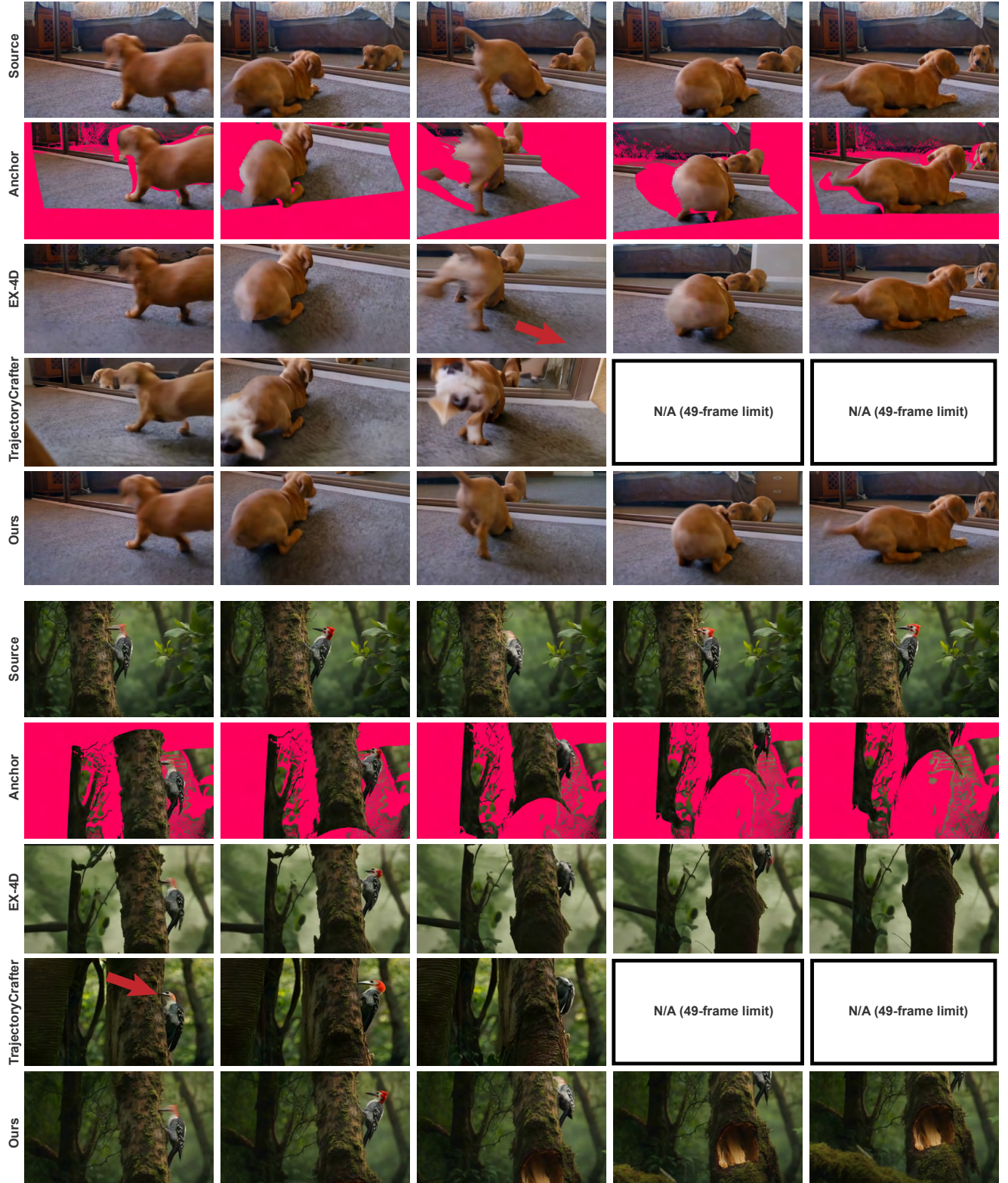


Figure 5. **Handling Random Camera Start Points while Maintaining Detail.** We illustrate anchor trajectories initiated from arbitrary viewpoints, causing significant spatial misalignment at the start of generation. Our model successfully handles these large initial viewpoint shifts without compromising quality, consistently preserving fine details and original textures.

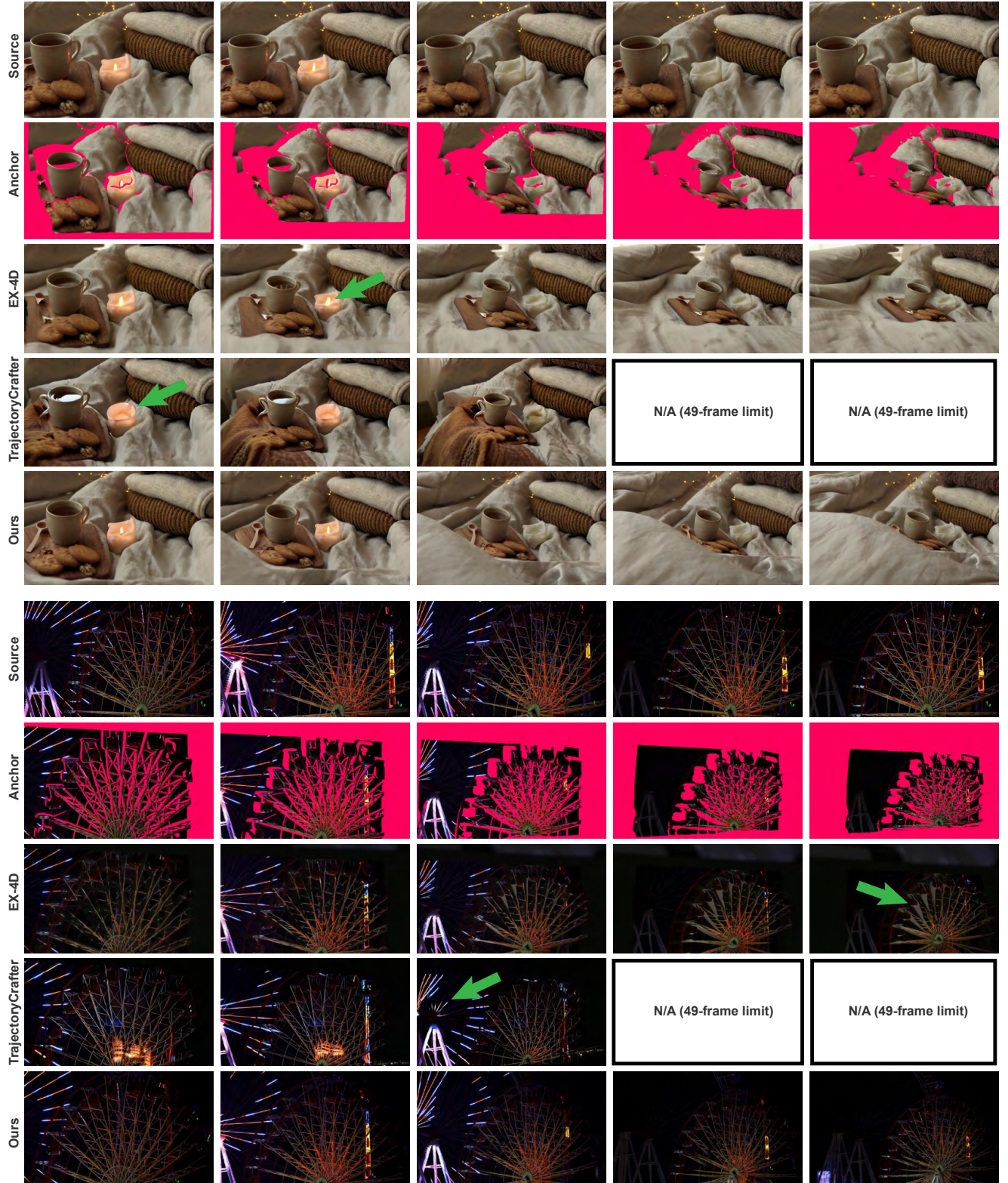


Figure 6. **Handling Random Camera Start Points while Maintaining Detail.** Additional examples of trajectories initiated from arbitrary viewpoints. Our model maintains exceptional stability and texture preservation under these challenging initialization conditions.

2.5. Extended Qualitative Comparisons and Applications

To provide a comprehensive visual analysis, we present extended qualitative comparisons in Figures 2, 3, and 4. Base-



Figure 7. **Extending Capabilities: Generative Outpainting.** Beyond standard video reshooting, our model demonstrates strong generative priors useful for outpainting tasks. In this example, a 2D crop window is shifted significantly towards the bottom-right, revealing a large unseen region. Crucially, unlike standard per-frame outpainting, our approach is full-context. The model attends to the entire input source video simultaneously, leveraging its learned implicit routing to plausibly hallucinate missing content and coherently complete the scene with high temporal stability.

line methods frequently exhibit characteristic artifacts, such as severe blurring, geometric distortion, or the loss of fine-grained details present in the source input. In contrast, our approach consistently demonstrates superior visual fidelity by effectively leveraging the stable scene priors provided by the source-video conditioning.

Furthermore, Figures 5 and 6 demonstrate the robustness of our model to challenging initialization conditions, enabled by our Random Query augmentation. Even when target trajectories initiate from random viewpoints causing massive spatial misalignment at the first frame, our model exhibits remarkable stability.

Finally, we highlight the versatility of our generative framework. Beyond standard reshooting, Figure 7 illustrates the model’s capacity for generative outpainting. When subjected to extreme 2D cropping that reveals previously unseen regions, the model does not simply hallucinate content frame-by-frame. Instead, it performs full-context outpainting. By attending to the entire source video sequence simultaneously via our conditioning architecture, the model can route available texture information from other temporal frames to plausibly and coherently fill the missing regions, ensuring strict temporal consistency.

References

- [1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 1
- [2] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 4
- [3] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025. 1
- [4] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2005–2015, 2025. 4
- [5] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025. 4
- [6] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 4
- [7] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaocong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhua Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 4
- [8] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024. 4
- [9] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 4
- [10] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1
- [11] Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6632–6644, 2025. 4