Unboxed: Geometrically and Temporally Consistent Video Outpainting Supplementary Material

Zhongrui Yu[1] Martina Megaro-Boldini[2] Robert W. Sumner[1,2] Abdelaziz Djelouah[2]

[1]ETH Zürich [2]DisneyResearch|Studios

zhonyu@ethz.ch abdelaziz.djelouah@disney.com

1. Computational Complexity

Table 1 compares the computational complexity of our method compared to M3DDM [2] and MOTIA [10]. We report the maximum GPU memory usage and the total runtime for both training and inference stage when outpainting the video from 284×480 to 854×480 as *Small FoV*(0.66 mask ratio, $3\times$ expansion) and from 480×480 to 2560×720 as *Large FoV* ($8\times$ expansion). To ensure fair comparison, we use a single 40GB V100 GPU for all methods.

The result shows that our method requires significantly less GPU memory than other approaches, particularly for large FoV outpainting, where M3DDM [2] encounters out-of-memory (OOM) and MOTIA [10] requires fine-tuning on a 40GB GPU but still fails to achieve satisfactory results. The majority time consumption of our method lies in the pre-processing stage, especially in estimating camera poses for video frames. For fine-tuning and inference, our method requires approximately 500 seconds more than MOTIA [10]. While M3DDM [2] requires the least time since it does not involve test-sample fine-tuning, it is worth noting that it underwent extensive training on a large-scale dataset beforehand (2.5 weeks on 24 A100 GPUs).

2. More Implementation Details

Static Region Reconstruction and Outpainting. In the first and last GS training rounds, GS point densification begins after 1000 iterations. In the intermediate training rounds that alternate with image outpainting, densification starts at 350 iteration. The learning rates for all GS parameters (position, opacity, color, rotation, scaling, *etc*.) remain unchanged from the original GS [3] settings. The coefficients for SSIM loss L_{SSIM}, depth lossL_{depth} and superpixel depth loss $L_{\text{depth}}^{\text{Sp}}$ are set at 0.2, 0.1, and 0.1, respectively. After outpainting each keyframe, the outpainted image is blended with the input image and added in the training set. The blending process employs a multi-band compositing strategy from [5] to ensure smooth transitions between the outpainted and original regions.

Figure 1. Steps for obtaining the masks for the dynamic regions: We use SAM [7] panoptic segmentation, then the epipolar error is used in a voting scheme to identify dynamic segments and obtain the final mask.

Initial Object Inpainting. During the fine-tuning of SDXL [6, 9] on the input sequence, the learning rates for the text encoder and the UNet are $4e-5$ and $2e-4$ respectively. The maximum training step is set to 2000, but in most cases, results after 800 steps are able to provide a good enough starting point for final video synthesis.

Mask Strategy. Mask accuracy is crucial for the static region outpainting step in our method, including the dynamic mask for reconstruction and the outpainting mask. Using accurate dynamic masks helps to reduce the artifacts during reconstruction. A straightforward approach to obtain these masks is to manually mask the region of interest in the first frame and use an object tracking model [8, 11] to track it across subsequent frames. Liu *et al.* [4] proposed an automatic way that estimates the epipolar error to identify the dynamic regions. However, this approach is sensitive to the threshold selection, which can result in incorrect or incomplete mask. A more effective approach involves using the Segment Anything Model 2 (SAM 2) [7] for panoptic segmentation of all video frames and then determine the dynamic regions by voting on the epipolar error within each segment, as shown in Fig. 1.

Accurate outpainting mask ensures coherence between the outpainted and original regions while preventing holes

	M3DDM		MOTIA		Ours		
	(a) Small FoV 3x expansion	(b) Large FoV 8x expansion	(a) Small FoV 3x expansion	(b) Large FoV 8x expansion		(a) Small FoV 3x expansion	(b) Large FoV 8x expansion
Fine Tuning (Sequence specific)	-	-	15.80 GB, 2068 s	36.78 GB, 4996 s	Pre-process	5.36GB, 3140s	
					Static region Outpainting	4.59 GB, 433 s	9.30 GB, 1765 s
					Object-specific Inpainting	11.30 GB, 2258 s	
Inference	17.95 GB, 431 s	OOM	7.15 GB, 155 s		Guided Video Synthesis	10.81 GB, 110 s	
In Toal	17.95 GB, 431 s	OOM	15.80 GB, 2223s	36.78 GB, 5151s		11.30 GB, 6384 s	11.30 GB, 7716 s

Table 1. Computation complexity of different methods in terms of peak GPU memory usage (GB) and running time (s). We use 2 different setting: (a) *Small FoV* $3\times$ *expansion* which corresponds to $284 \times 480 \rightarrow 854 \times 480$, and (b) *Large FoV* $8\times$ *expansion* which corresponds to $480 \times 480 \rightarrow 2560 \times 720$.

in outpainting. Previous approach [1] renders the opacity from the GS model and defines the mask using a threshold on each pixel's opacity value. However, the artifacts introduced in GS training often lead to poor mask quality. To address this, we create an additional isotropic 3D Gaussian ball for each GS model point, splat them onto the image plane, and mask out the pixels that are not covered. This approach ensures a more accurate boundary between the existing region and the region to be outpainted, enabling consistent and seamless outpainting, as shown in Fig. 3.

3. Visual Results for Ablation Studies

Figure 2 shows the visual results of the ablation study in the main paper. In Fig. 2 (a), we show the results of frame-by-frame outpainting using the fine-tuned image model [6, 9] on the input video frames. While the outpainted regions are visually plausible and faithful to the orignal content, the temporal consistency is not guaranteed. Fig. 2 (b) shows the guided video synthesis results built upon results from (a). The temporal consistency is improved, as observed in the flowerpot on the left side of the frames, showcasing the potential of the guided video synthesis process. However, its effectiveness is limited by the quality of the starting point. In Fig. 2 (c) and (d), the consistency of the static background is ensured by our GS supported outpainting. However, the object appearance varies over time without the object-specific inpainting or the guided video synthesis in our pipeline. Fig. 2 (e) demonstrates that our complete model achieves the best results.

4. More Details on User Study

Figure 4 shows the interface designed for our user study. In the study, participants were presented with the input video, followed by the outpainting results generated by three methods (M3DDM[2], MOTIA [10], and ours) displayed in random order.

Figure 2. Qualitative ablation results on different modules in our method. We draw the dynamic object bounding box (in green) when this concept is used. (a). Outpainting with fine-tuned Image Model; (b). (a) + Guided Video Synthesis; (c). Our model without guided video synthesis; (d). Our model without the initialization from object inpainting; (e). Our complete model.

Figure 3. Obtaining the outpainting mask for the static scene outpainting and reconstruction step.

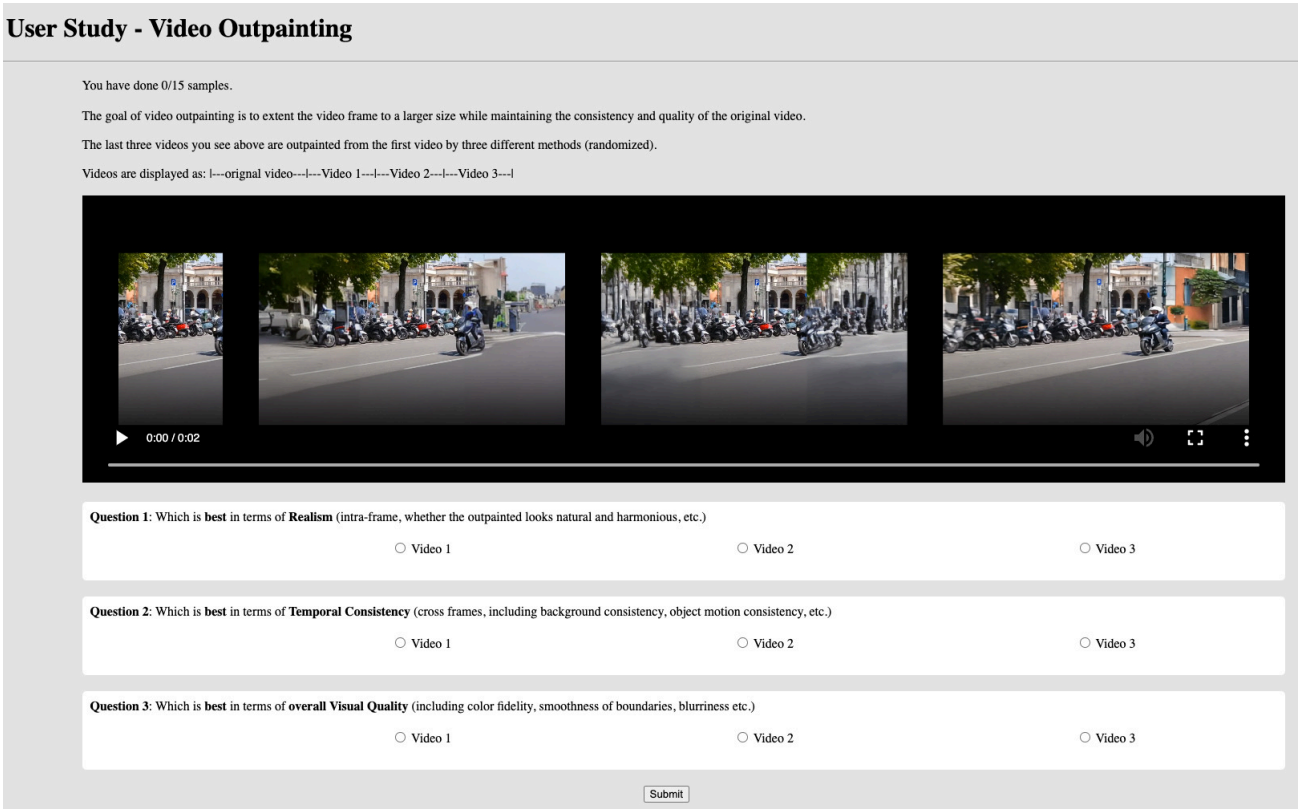

Figure 4. Screenshot of our user study interface.

References

[1] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 2

[2] Fanda Fan, Chaoxu Guo, Litong Gong, Biao Wang, Tiezheng Ge, Yuning Jiang, Chunjie Luo, and Jianfeng Zhan. Hierarchical masked 3d diffusion model for video outpainting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7890–7900, 2023. 1, 2

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[4] Yu-Lun Liu, Chen Gao, Andréas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13–23, 2023. 1

[5] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, pages 173–184. Wiley Online Library, 2020. 1

[6] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[8] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 1

[9] 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, 2

[10] Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. *arXiv preprint arXiv:2403.13745*, 2024. 1, 2

[11] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 1