
Supplementary Material of Instant4D: 4D Gaussian Splatting in Minutes

Anonymous Author(s)

Affiliation

Address

email

1 Additional Evaluation

2 **Qualitative comparison on the DAVIS** We present qualitative comparisons with the concurrent
3 method RoDyGS [3] on the DAVIS [10] dataset, as shown in Figure 1. Our method accurately captures
4 the motion of the bear, particularly in challenging regions such as the torso and feet. Throughout the
5 sequence, our method preserves details and temporal consistency in both small, fast-moving regions
6 (e.g., the feet) and large deformable structures (e.g., the torso), while RoDyGS frequently introduces
7 blurring and artifacts in these areas. For instance, at Frame 40, RoDyGS renders the bear’s feet with
8 transparency, whereas our method preserves the structural integrity and appearance of these limbs.
9 Furthermore, our approach faithfully reconstructs fine-grained details, such as the texture of the bear’s
10 fur, while the rendering of RoDyGS appears over-smoothed or missing. The visual comparison for
11 the Bear scene highlights the robustness of our approach in handling motion and preserving details.

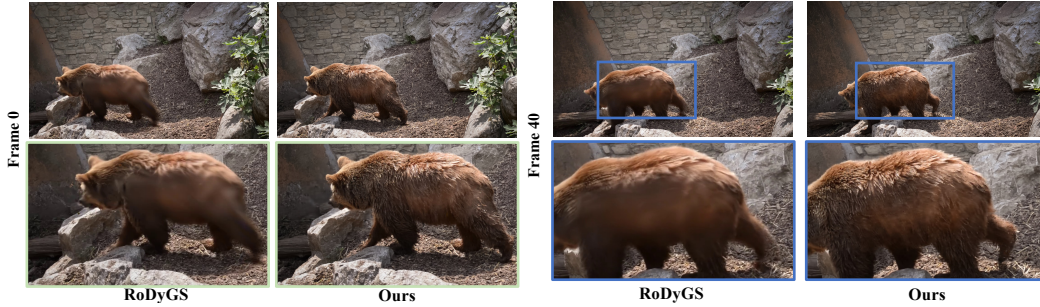


Figure 1: Visualization of the Bear scene in the DAVIS [10] dataset.

12 As shown in Figure 2, our method accurately reconstructs the rhino’s skin with sharp texture and
13 rich shading detail, effectively capturing the light and surface geometry. In contrast to RoDyGS [3],
14 which introduces artifacts as motion blurs and a loss of structure, our method preserves both spatial
15 details and temporal consistency. Notably, the boundaries of the reconstructed scene remain clean,
16 with few ghosting or artifacts outside the viewing frustum. We argue that it is attributed to the usage
17 of back-projected point clouds as initialization combined with isotropic Gaussian primitives, which
18 together help to constrain geometry and appearance within the observable volume.

19 **Discussion with previous work** Our first goal is to decouple geometric recovery from photometric
20 optimization. RoDynRF [7] back-propagates reprojection error through a static radiance field to
21 refine camera poses, a process that exceeds 24 hours (h) per scene. InstantSplat [1] likewise optimize
22 camera extrinsics during training, but the joint optimization of Gaussians and poses increases runtime
23 and introduce a position-pose ambiguity. In contrast, our model reconstructs a scene in **0.03h**, and
24 outperforms RoDynRF by **6.22dB** on the Dycheck dataset [2] and InstantSplat by **1.43dB** (PSNR) on
25 the NVIDIA dataset [16].

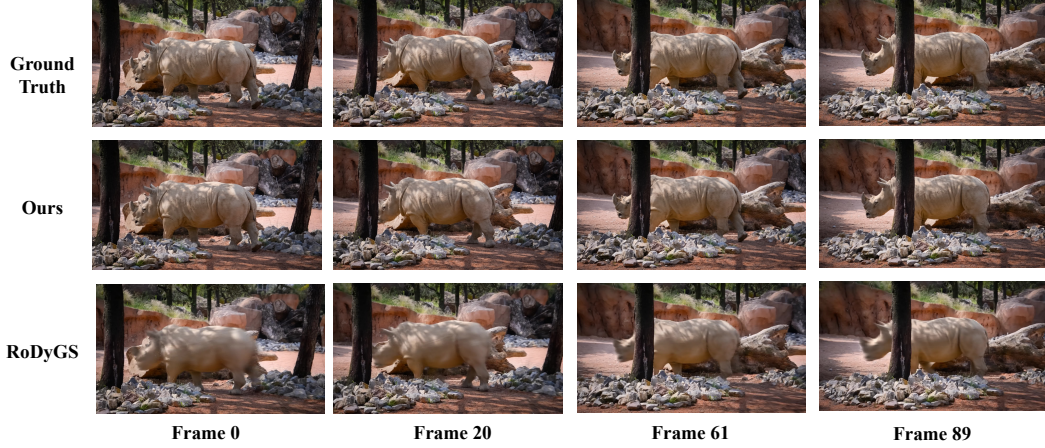


Figure 2: Visualization of the Rhino scene in the DAVIS [10] dataset.

Because pose refinement is handled separated from optimization, our method needs only a single photometric loss. By comparison, recent Gaussian-based pipelines introduce rigidity [12, 4, 13], point track [4, 13, 12], or depth regularization [13, 4] to stabilize training and better estimate cameras. Even without these auxiliary terms, we match or exceed state-of-the-art accuracy on both the NVIDIA Dynamic Scene and Dycheck iPhone datasets.

Our second goal is to eliminate redundant primitives while preserving occlusion structure. RoDyGS [3] encodes motion in a shallow MLP based on the MAST3R [5], redundantly storing identical background content that re-appears across frames. Leveraging grid-pruned, motion-aware 4D Gaussians [15] removes such duplication: we are **20×** faster than RoDyGS and achieve a **7.15dB** PSNR gain. Comparing with 4DGS baseline, grid pruning leads to an **8×** acceleration on the NVIDIA dataset while improving visual quality.

Table 1: Breakdown of Runtime.

Component	Runtime (Sec)	Memory (M)
Geometry Recovery		
Depth Estimation	23.98	2935
Camera Tracking	23.47	9602
Video Depth Optimization	98.55	5501
Grid Pruning	2.76	-
Optimization		
Forward Splatting	55.47	-
Backward	92.53	3878
Total Training Time	295.76	9602

2 Methods Detail& Ablation Analysis

2.1 Runtime Analysis

Table 1 reports a fine-grained runtime and memory profile of the entire pipeline. The experiment is conducted on a single A6000 GPU with an 82-frame input video of 854×480 resolution. We report the breakdown of our model’s running time. The whole end-to-end training finishes in 6 minutes after which the model renders at 247 FPS in real time. The proposed grid pruning only takes less than 3 seconds and overall optimization process only takes less than 4G memory, which showcase that our design is less redundant and light-weighted.

2.2 Additional Discussion on Motion Awareness

We provide additional analysis on the how we get the motion mask from motion probability in our visual SLAM process. As stated in the Section 3.1 in our paper, MegaSAM [6] maintains a per-frame disparity map $\hat{\mathbf{d}}_i \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8}}$ as well as a motion probability prediction $\hat{m}_i \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8}}$ in order to calculate the reprojection error for the static region. We interpolate \hat{m}_i into original resolution and get $m_i \in \mathbb{R}^{H \times W}$. After that, we employ Otsu’s method[9] to obtain a binary mask for the motion segmentation. Next, we assign temporal scaling based on this segmentation. As illustrated in the Figure 3, after adding pseudo-frame, the segmentation get rid of the noise due to movement of camera e.g. at the left of the picture. Besides, this strategy also reduce the dynamic region area and therefore reduce the computation overhead. By comparing with the ground truth motion mask, we find that our motion mask looks eroded by few pixels, which is caused from the interpolation from \hat{m}_i to m_i .

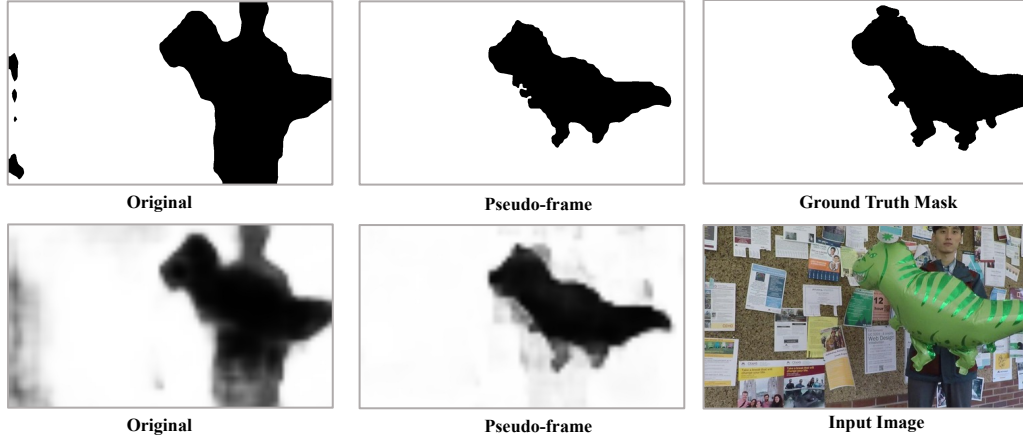


Figure 3: Visualization for the our binary motion mask from predicted motion probability.

Our proposed method is simple, effective and fully automated without human annotation. In contrast, other work employ human interaction to segment object [13], apply extra tracking model [14] to get the moving object bounding box [3] or calculate flow error map [17] as annotation for Segment Anything model [11] in order to generate the motion mask.

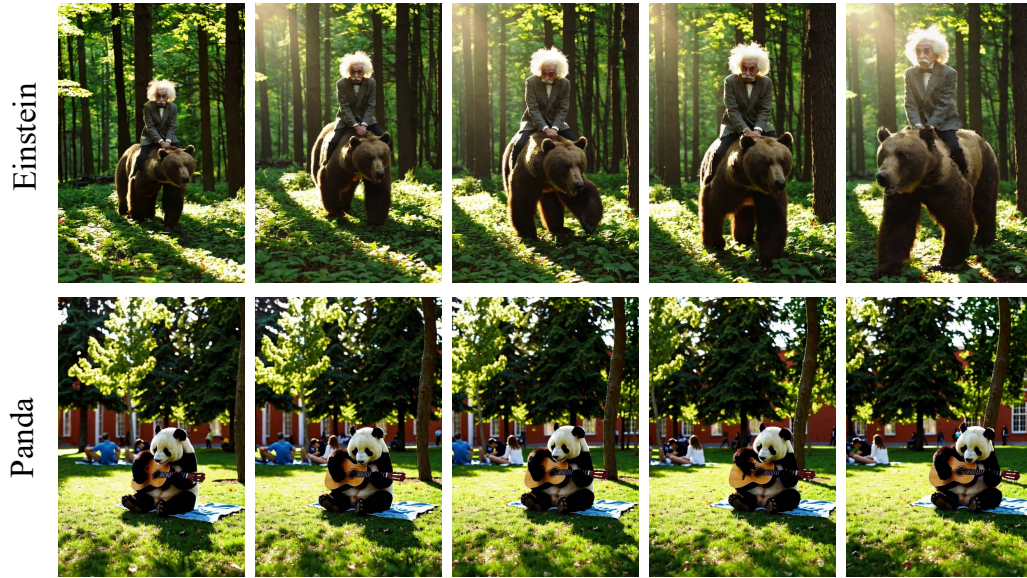


Figure 4: Reconstruction for Sora [8] generated Video

60 **3 Visualization on the AIGC Video**

61 Recent advances in generative models, such as Sora [8], enable the synthesis of photorealistic videos
62 with dynamic camera motion and complex scenes. One important application of our method is
63 integrating with the AI generated content (AIGC) creation. We apply our reconstruction pipeline
64 to Sora generated video. As illustrated in Figure 4, we generate a 5-second video using Sora with
65 prompts such as Einstein riding a bear and a panda playing guitar. More results are included in the
66 supplementary files.

References

- [1] Zhiwen Fan, Kairun Wen, Wenyan Cong, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Sparse-view sfm-free gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024.
- [2] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Dynamic novel-view synthesis: A reality check. In *NeurIPS*, 2022.
- [3] Yoonwoo Jeong, Junmyeong Lee, Hoseung Choi, and Minsu Cho. Rodygs: Robust dynamic gaussian splatting for casual videos. *arXiv preprint arXiv:2412.03077*, 2024.
- [4] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024.
- [5] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.
- [6] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024.
- [7] Yu-Lun Liu, Chen Gao, Andreas 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*, pages 13–23, 2023.
- [8] OpenAI. Sora: Creating videos from text. <https://openai.com/sora>, 2024. Accessed: 2025-05-22.
- [9] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.
- [10] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.
- [11] 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.
- [12] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.
- [13] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.
- [14] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.
- [15] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023.
- [16] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020.
- [17] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.