
Supplementary Material for Motion4D: Learning 3D-Consistent Motion and Semantics for 4D Scene Understanding

In this supplementary material, we provide additional information to complement the main paper. First, we present the implementation details and experimental settings of Motion4D (Section A). Second, we provide further information about the proposed DyCheck-VOS benchmark (Section B). Third, we include additional experimental results on the DyCheck dataset (Section C). Lastly, we present qualitative visualizations on the DAVIS dataset (Section D).

A Implementation Details

In this section, we provide full implementation details of Motion4D.

Data processing. We leverage off-the-shelf 2D models to obtain prior knowledge for initializing the 4DGS model. For DyCheck dataset [2], we use COLMAP [12] camera poses and metric depth maps processed by [13]. For in-the-wild videos from the DAVIS dataset [10], we adopt MegaSaM [7] to estimate per-frame camera poses and depth maps. In all the experiments, we use SAM2 [11] with first-frame masks as prompts to generate initial segmentation inputs. We then use TAPIR [1] to compute long-range 2D tracks by querying points sampled from the foreground masks and tracking them across all frames.

Initialization. We follow a similar paradigm to [13] for initializing the 4DGS model. To initialize the dynamic part of the scene, we first identify the foreground objects using segmentation masks and sample 2D tracks within the masked regions. The 2D tracks are lifted into 3D using the estimated depth maps, and the resulting 3D coordinates in the canonical frame are used to initialize the means of the dynamic Gaussians. Then, for the motion field, we first apply k-means clustering to the Gaussian means, resulting in B clusters of tracks, each corresponding to an initial motion basis. Each motion basis is initialized by solving a Procrustes problem between the corresponding 3D point sets at different time steps, resulting in an estimated $\mathbb{SE}(3)$ transformation. For the static background, we use standard static 3D Gaussians, with their positions initialized by sampling points from the depth maps. The static and dynamic Gaussians are jointly optimized and rasterized together to form an image.

Iterative refinement. The adaptive resampling module is designed to explicitly insert new Gaussians into the underrepresented regions based on photometric error or semantic error. At the start of each iteration, we sample a maximum of 4,000 points from all frames. In the iterative semantic refinement process, we introduce up to 5 additional prompts per object, with new prompts added at an interval of 30 frames.

Optimization. We optimize our model using Adam Optimizer [5]. We train 30 epochs for every sequence during sequential optimization with a sequence length of 30 frames. After sequential optimization on both motion and semantic fields, we perform 200 epochs on global optimization for training all the fields jointly. We perform the same adaptive Gaussian controls for dynamic and static Gaussians as 3DGS [4]. All experiments are conducted on a single NVIDIA 4090 GPU.

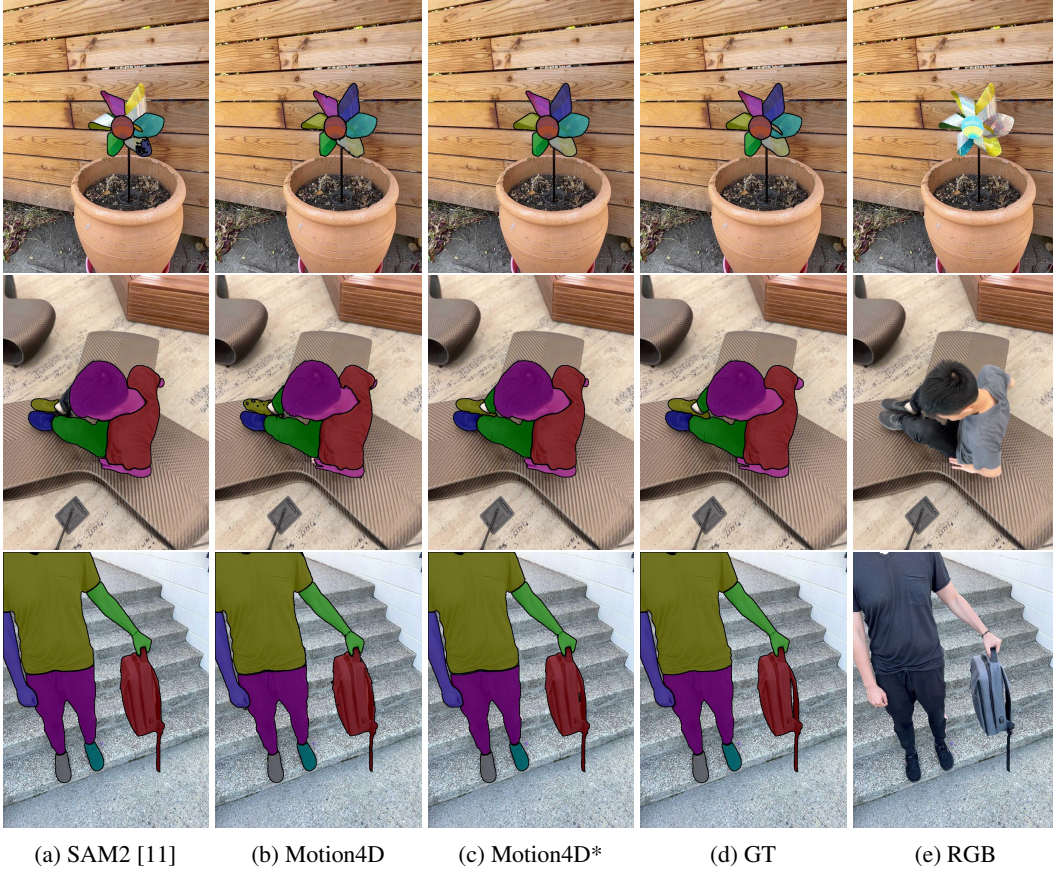


Figure 1: Visualization of segmentation results on the DyCheck-VOS benchmark. We provide our results of both rendered masks (Motion4D) and refined SAM2 masks (Motion4D*).

Table 1: Comparison of point-based tracking and novel view synthesis results on DyCheck [2] dataset.

Method	3D Tracking			2D Tracking			Novel View Synthesis		
	EPE ↓	$\delta_{3D}^{05} \uparrow$	$\delta_{3D}^{10} \uparrow$	AJ ↑	$< \delta_{avg} \uparrow$	OA ↑	PSNR ↑	SSIM ↑	LPIPS ↓
T-NeRF [2]	-	-	-	-	-	-	15.60	0.55	0.55
HyperNeRF [9]	0.182	28.4	45.8	10.1	19.3	52.0	15.99	0.59	0.51
DynIBaR [6]	0.252	11.4	24.6	5.4	8.7	37.7	13.41	0.48	0.55
Deformable-3D-GS [15]	0.151	33.4	55.3	14.0	20.9	63.9	11.92	0.49	0.66
CoTracker [3]+DA [14]	0.202	34.3	57.9	24.1	33.9	73.0	-	-	-
TAPIR [1]+DA [14]	0.114	38.1	63.2	27.8	41.5	67.4	-	-	-
Shape of Motion [13]	0.082	43.0	73.3	34.4	47.0	86.6	16.72	0.63	0.45
Motion4D	0.072	46.7	75.9	37.3	50.4	87.1	17.91	0.69	0.42

B DyCheck-VOS benchmark

We introduce the DyCheck-VOS benchmark to evaluate video object segmentation performance in realistic and dynamic scenes. DyCheck-VOS consists of 14 sequences, each spanning 200–500 frames, with 2–7 annotated foreground objects per sequence. The video resolution is 720×960 . We first apply SAM2 [11] to generate initial per-frame object masks, and manually refine the masks using brush-based annotation tools. As shown in Figure 1, our annotations offer high-quality masks that capture fine details and challenging partial regions, such as hands, clothing, and other small or disconnected parts. We find that 2D models often struggle to produce consistent results, particularly for partially visible objects, where a deeper 3D understanding is required.



Figure 2: Visualization of segmentation results on the DyCheck-VOS benchmark. We provide our results of both rendered masks (Motion4D) and refined SAM2 masks (Motion4D*).

C Additional Results on the DyCheck Dataset

Novel view synthesis results. We report quantitative results of novel view synthesis in Table 1. We compare Motion4D with state-of-the-art Nerf [8] and 3DGS [4] methods, which demonstrates that our method consistently outperforms all baselines by a substantial margin. Figure 2 provides novel view synthesis comparison on the validation views. The regions highlighted in green indicate areas excluded from evaluation due to the lack of co-visibility between training and validation views.

3D tracking results. We also evaluate the performance of 3D point tracking in Table 1. Following [13], we generate ground truth 3D tracks by projecting the 2D keypoint annotations into 3D using lidar depth. We then evaluate the tracking performance and report the 3D end-point-error (EPE) and the percentage of points that fall within a given threshold $\delta_{3D}^{0.5} = 5\text{cm}$ and $\delta_{3D}^{1.0} = 10\text{cm}$. We show that Motion4D outperforms all previous methods in terms of 3D scene motion estimation.

D Additional Results on the DAVIS dataset

Finally, in Figure 3, we show view synthesis results of training views of in-the-wild videos from the DAVIS dataset [10]. Motion4D is capable of rendering high-quality views, which demonstrates accurate reconstruction quality and provides reliable scene geometry for estimating semantics and motion.

References

- [1] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023.
- [2] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022.
- [3] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024.
- [4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- [5] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

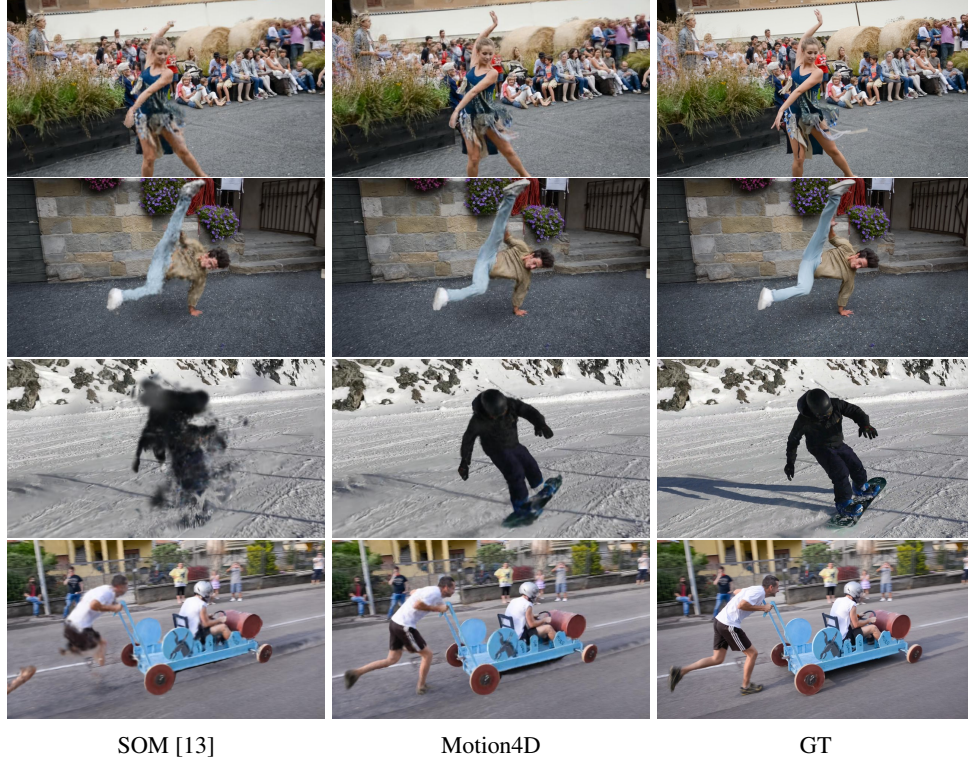


Figure 3: Visualization of rendered views on the DAVIS dataset.

- [6] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023.
- [7] 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.
- [8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [9] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.
- [10] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 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] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [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] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

- [15] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024.