MultiEgo: A Multi-View Egocentric Video Dataset for 4D Scene Reconstruction

Bate Li
Shanghai Jiao Tong University
Shanghai, China
woxelikeloud@sjtu.edu.cn

Houqiang Zhong
Shanghai Jiao Tong University
Shanghai, China
zhonghouqiang@sjtu.edu.cn

Zhengxue Cheng*
Shanghai Jiao Tong University
Shanghai, China
zxcheng@sjtu.edu.cn

Qiang Hu
Shanghai Jiao Tong University
Shanghai, China
qiang.hu@sjtu.edu.cn

Qiang Wang
VisionStar Information Technology (Shanghai) Co., Ltd.
Shanghai, China
wq@sightp.com

Li Song*
Shanghai Jiao Tong University
Shanghai, China
song_li@sjtu.edu.cn

Wenjun Zhang*
Shanghai Jiao Tong University
Shanghai, China
zhangwenjun@sjtu.edu.cn

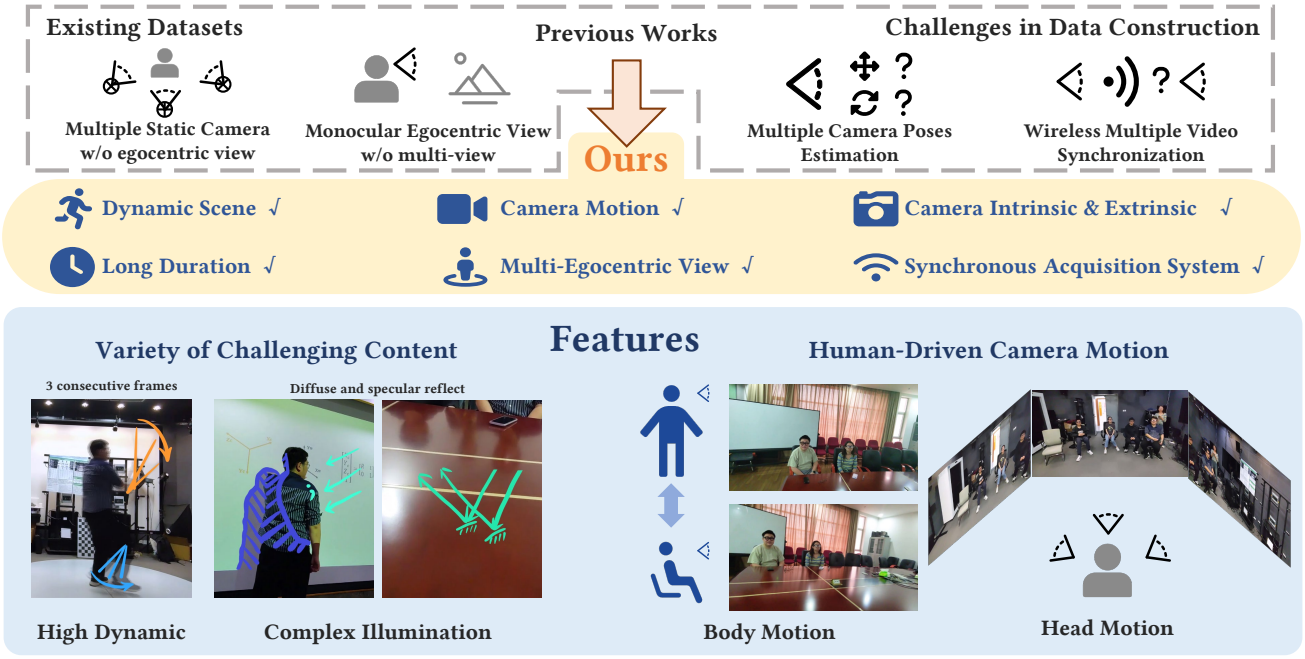

Figure 1: MultiEgo is the first multi-egocentric dynamic scene reconstruction dataset. The dataset provides essential data for reconstruction tasks, including synchronized egocentric videos and accurate camera pose annotations. It also features various challenges for reconstruction, such as human-driven camera motion, high-dynamic objects, and complex illumination.

*Corresponding authors.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758232

Abstract

Multi-view egocentric dynamic scene reconstruction holds significant research value for applications in holographic documentation of social interactions. However, existing reconstruction datasets focus on static multi-view or single-egocentric view setups, lacking multi-view egocentric datasets for dynamic scene reconstruction. Therefore, we present MultiEgo, the **first** multi-view egocentric dataset for 4D dynamic scene reconstruction. The dataset comprises five canonical social interaction scenes: meetings, performances, and a presentation. Each scene provides five authentic egocentric videos captured by participants wearing AR glasses. We design a

Table 1: Comparison of MultiEgo and existing egocentric and 4D reconstruction Datasets.

Character	Egocentric Datasets			4D reconstruction Datasets			Ours
	Ego4D [13]	Ego-Exo4D [14]	Epic-Kitchens [6–8]	N3DV [25]	D-NeRF [39]	HyperNeRF [36]	
Dynamic scene	✓	✓	✓	✓	✓	✓	✓
Egocentric view	✓	✓	✓	✗	✓	✓	✓
Multi-perspective	✓	✓	✗	✓	✗	✗	✓
Multi-egocentric view	✓	✗	✗	✗	✗	✗	✓
Camera poses provided	✗	✓	✗	✓	✓	✓	✓

hardware-based data acquisition system and processing pipeline, achieving sub-millisecond temporal synchronization across views, coupled with accurate pose annotations. Experiment validation demonstrates the practical utility and effectiveness of our dataset for free-viewpoint video (FVV) applications, establishing MultiEgo as a foundational resource for advancing multi-view egocentric dynamic scene reconstruction research. Our project page and dataset are available at https://woxelikeloud.github.io/multiego/.

CCS Concepts

• **Computing methodologies** → **Computer graphics**; **Computer vision**.

Keywords

Egocentric Video; Free-viewpoint Video; Dynamic Scene

ACM Reference Format:

Bate Li, Houqiang Zhong, Zhengxue Cheng, Qiang Hu, Qiang Wang, Li Song, and Wenjun Zhang. 2025. MultiEgo: A Multi-View Egocentric Video Dataset for 4D Scene Reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758232

1 Introduction

Free-viewpoint video (FVV), powered by dynamic scene modeling [3, 5, 10, 16, 18, 25, 36, 55, 56], represents a transformative leap in next-generation visual representation through its capacity for immersive real-time interaction. This paradigm unlocks unprecedented opportunities in entertainment, virtual reality, and human-computer interaction. Recent advances in novel view synthesis, such as 3D Gaussian Splatting [22] (3DGS) have propelled dynamic scene reconstruction to new heights of efficiency and fidelity [17, 28, 33, 44, 49–51]. Meanwhile, the rise of smart wearable devices (e.g. AR glasses) is redefining acquisition paradigms through lightweight, egocentric capture [6, 7, 11, 14, 31, 35, 45]. Unlike conventional fixed multi-camera systems [4, 24, 25, 40], multi-user wearable frameworks offer dual advantages: 1) eliminating interference of acquisition equipment while preserving natural participant behaviors, 2) enabling 4D social scene reconstruction via multi-view fusion [20, 21, 43] in a more convenient way. These innovations pave the way for practical applications such as FVV meeting summaries and holographic concert recordings, heralding a new era of deployable dynamic scene reconstruction.

However, existing datasets often exhibit critical limitations in multi-egocentric view reconstruction research: (1) Most dynamic scene datasets such as N3DV [25], employ fixed multi-camera setups with static viewpoints, whereas multi-egocentric capture is driven by natural human-driven motions; (2) While monocular moving viewpoint datasets like HyperNeRF [36] and D-NeRF [39], lack multiple egocentric views, resulting in less information in reconstructing; (3) Egocentric-centric datasets such as HOI4D [31], EPIC-Kitchens [6–8], EgoDex [15] primarily focus on human-object interactions or activity recognition tasks, including comprehensive benchmarks like Ego4D [13] and Ego-Exo4D [14] which emphasize video understanding tasks rather than scene reconstruction. In a nut shell, no existing dataset simultaneously provides multi-person egocentric perspectives with synchronized pose estimations, which is the critical capability our dataset explicitly addresses.

In this paper, we present MultiEgo, the **first** multi-egocentric dynamic scene reconstruction dataset, addressing the limitations of existing datasets that primarily focus on fixed-camera and monocular egocentric settings. The dataset contains five dynamic scenes, including meetings, performances, and a presentation. Each was captured through five 1080p 30 FPS egocentric videos with accurate pose annotations. By making full use of the device, we addressed the synchronization challenges during recording. We conducted baseline evaluations and detailed analysis that validated the effectiveness of the dataset and provide insights for future task development. Our dataset characteristics are compared with existing datasets, including egocentric datasets: Ego4D [13], Ego-Exo4D [14], EPIC-Kitchens [6–8], and 4D reconstruction datasets: N3DV [25], D-NeRF [39], HyperNeRF [36], as summarized in Table 1.

Our contribution could be summarized as follows:

- We present the MultiEgo dataset, the first dataset for multi-egocentric dynamic scene reconstruction. The dataset contains five challenging daily social scenes, and each scene is composed of five strictly synchronized egocentric videos with accurate pose annotations.
- We design a customized multi-egocentric data acquisition system and a data process pipeline, enabling hardware-level synchronization across viewpoints and accurate pose estimation.
- We conducted baseline evaluations on our dataset. The experimental results demonstrate its strong practical utility and effectiveness for dynamic scene reconstruction tasks.

2 Related Works

Dynamic Scene Reconstruction Dataset. With the advancement of computer vision technologies and data acquisition devices, numerous dynamic scene datasets have emerged [4, 24, 25, 39, 40, 52]..

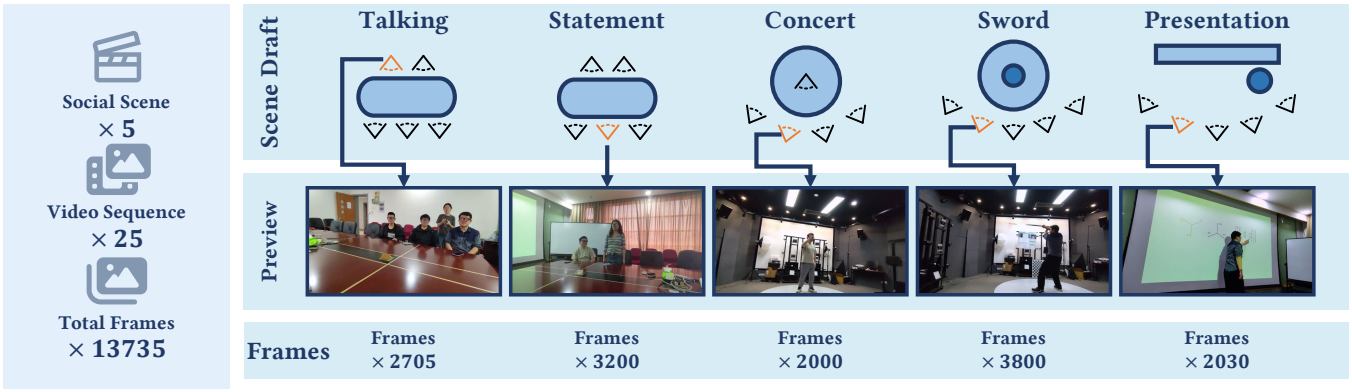

Figure 2: An overview of our MultiEgo dataset. There has 5 high-quality dynamic scenes in the dataset, each scene contains 5 egocentric views with accurate pose annotation. The total number of frames comes to 13,735.

Representative dynamic datasets such as the N3DV [25] dataset provides multi-view video recordings from fixed perspectives to capture dynamic scenes, where one or more performers perform various subtle interactions, such as cooking and talking. However, perspective coverage in N3DV is limited, and fixed viewpoints require specialized equipment for similar data collection. The HyperNeRF [36] dataset features scenes captured from a single moving viewpoint, recording brief actions such as breaking cookies or pouring liquids. Although data in HyperNeRF can be considered egocentric, it focuses on small-scale object-centric scenarios with only monocular observations.

Egocentric Dataset. In recent years, egocentric vision datasets have experienced rapid development [2, 6–9, 11, 13–15, 23, 26, 31, 32, 34, 35, 42, 45]. Prominent examples include Ego4D [13] and EgoExo4D [14], which feature large-scale data volume and diverse content modalities. However, these datasets typically provide only single-view egocentric perspectives and are primarily designed for video content understanding [27, 37, 46] rather than scene reconstruction. With the emergence of multi-modal large models, several human-object interaction (HOI) datasets have been proposed, including HOI4D [31], Epic-Kitchens [6–8], and EgoDex [15]. While these datasets advance tasks like human behavior understanding, their design objectives and data characteristics make them unsuitable for direct application in dynamic scene reconstruction tasks. For example, EgoGaussian [53] attempts to reconstruct egocentric datasets like HOI4D [31] but still requires excessively complex pre-processing, such as performing hand-object segmentation on frames and then estimating camera poses using COLMAP [41].

3 MultiEgo Dataset

3.1 Scene Overview

To address the limitation that existing datasets are inapplicable to multi-view egocentric dynamic scene reconstruction, we present the MultiEgo dataset, the **first** dataset for multi-egocentric dynamic scene reconstruction. The dataset contains five multi-person social scenes, including meetings, performances, and a presentation. Each scene has five performers wearing AR glasses to provide authentic and reliable egocentric views. All participants signed informed consent forms authorizing the use of facial and other biometric data for academic research purposes.

The first scene called ***talking*** involves a discussion meeting, in which performers take turns speaking. During this structured interaction, the performers tend to focus on the active speaker, resulting in systematic rotational camera movements through natural head rotations. This patterned motion enabled the system to capture comprehensive scene coverage despite the limited field-of-view of the individual cameras. Notably, although only five cameras were deployed, human-driven camera motions introduced diverse motion patterns that effectively enriched the observed information. In addition, the smooth wall and table surfaces of the meeting room induce extensive specular reflections from artificial lighting sources, accompanied by enhanced diffuse environmental reflections. This illumination phenomenon imposes significant challenges on reconstruction methods in determining the spectral properties of the surface reflectance while maintaining photometric consistency under complex lighting conditions.

The second scene called ***statement*** involves a statement meeting, in which the performers take turns giving speeches while being required to stand up during their turns. During the transition from sitting to standing, we observed rapid translational movements caused by vertical body movement. In addition, after assuming the standing posture, certain performers exhibited natural exploratory behaviors characterized by horizontal head rotations, thereby enriching multi-perspective scene representations. Moreover, due to occlusion constraints in seated postures where performers cannot perceive others' lower body regions, the standing-up action effectively introduces new content in the scene, analogous to liquid pouring. This phenomenon imposes challenges on reconstruction algorithms.

The third scene called ***concert*** features a standing performance paradigm with an actor and four audience members, all equipped with AR glasses. This configuration simulates real-world standing events such as concerts or public speeches, enabling dual-perspective analysis: (1) reconstructing actor's embodied performance from audience-centric viewpoints, and (2) capturing audience reactions through the actor's egocentric viewpoint. Although the actor constitutes the sole active visual source directed towards

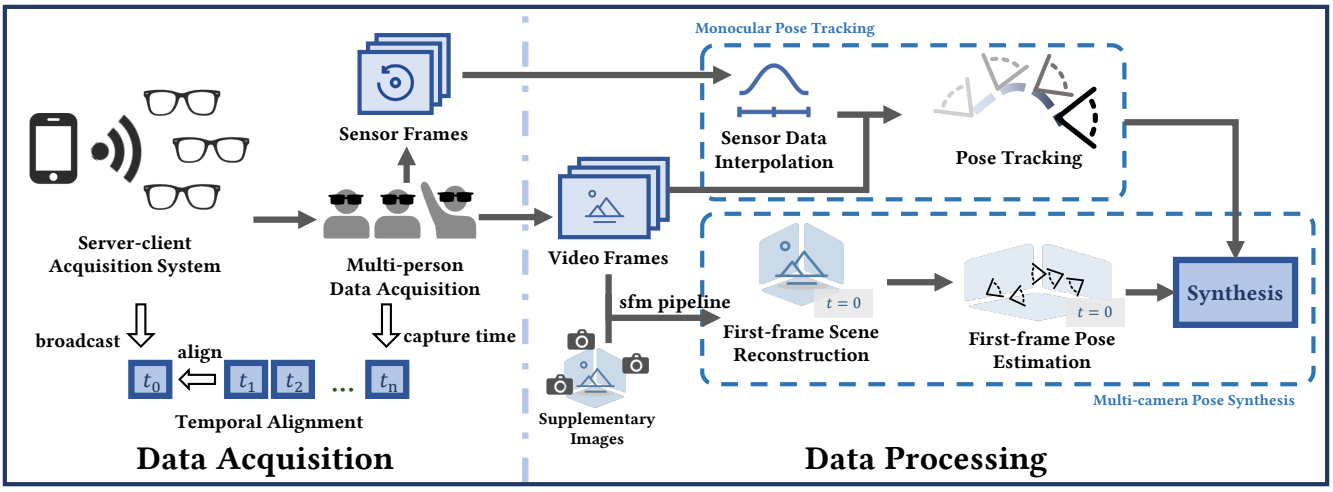

Figure 3: Pipeline of data acquisition and processing. We use a server-client system to handle synchronization, and respectively process time, monocular pose, poses of multiple views in the first frame. Finally we obtain the images and pose tracking in 4D domain.

the audience, the frequent head/upper-body movements during performance generate dense spatial-temporal sampling of the scene, while the audience's motion magnitude remains small due to their stationary postures. This introduces valuable priors for novel view synthesis.

The fourth scene called ***sword*** focuses on the dynamic action performance, where five audience members equipped with AR glasses observe an actor performing sword-wielding demonstrations using a long sword prop. This setup features sustained high-dynamic content characterized by complex 3D motion patterns: the limb movements of the actor exhibit rapid translational velocities, while the prop generates high angular velocities during slashing motions. These extreme motion characteristics pose significant challenges for high-speed motion reconstruction capabilities.

The fifth scene called ***presentation*** simulates educational settings, for example lecture recordings, through a slide presentation task, in which five audience members equipped with the glasses observe an actor executing a presentation and delivering exaggerated gestures. This setup mimics real-world projection-based presentations by capturing two critical visual phenomena: (1) dynamic occlusion shadows formed when the performer's body blocks the projected light path and (2) illumination-induced chromatic variations that emerge on the performer's body as they move near the projected screen. The latter effect arises from the complex interplay between ambient lighting and high-luminance projector beams.

Each of the five scenes presents distinct technical challenges, including rapid object motions, high-speed camera movements, and specialized illumination conditions, such as specular surface reflections. These challenges represent critical prerequisites that must be addressed before advancing multi-egocentric reconstruction toward broad applications. In order for our dataset to enable rigorous evaluation of algorithm robustness under these challenging conditions, it explicitly incorporates scenarios specifically designed to assess resilience against complex visual phenomena.

3.2 Data Acquisition

Hardware. In this paper, we select RayNeo X2 AR glasses for data collection, which is a consumer-grade AR device. The RayNeo X2 is equipped with a camera capable of recording 1080P resolution video at 30 frames per second. The device runs an Android operating system and supports WiFi communication. The official SDK provides real-time 3-degree-of-freedom rotational pose estimation from the built-in gyroscope sensor, which serves as an important foundation for camera pose estimation.

Acquisition System. To fully exploit the capabilities of the device, we developed a dedicated application system for data acquisition. The system integrates the functions of video capture, synchronized control, and sensor data collection. The system architecture adopts a client-server model: the client program runs on the AR glasses, continuously monitoring the WiFi channel to detect record/stop commands from the server. The server program runs on an external smartphone, establishing connections with the clients through a WiFi hotspot, and employs broadcast signals to synchronize recording start/end operations across multiple devices.

Once the server program broadcasts the start signal, all clients initiate data acquisition within several microseconds, which can be regarded as practically simultaneous activation. This level of synchronization aligns with the requirements for multi-perspective scene reconstruction. During data acquisition, the client simultaneously captures visual and sensor data streams. The camera records 1080P video at 30 frames per second while the gyroscope provides rotation outputs at 50 Hz sampling rate. Notably, the client program records timestamps for each data frame (both video and sensor) in UTC with 100-nanosecond precision during collection, ensuring temporal synchronization across modalities and devices.

Table 2: The quantitative results of our validation experiments. Bold: Best result. Underline : Second-best result.

Method	Talking			Statement			Concert		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGStream [44]	22.1335	0.7262	0.3382	20.4199	0.6711	0.3974	21.8715	0.7551	0.3074
Deformable-3DGS [50]	23.2186	0.8023	0.3358	21.3670	0.7731	0.3819	24.1020	0.8418	0.2886
4DGaussian [49]	**24.9863**	0.8094	0.3353	**24.0491**	**0.7894**	**0.3672**	**25.9235**	**0.8512**	0.2953
4DGaussian [49] w/ timestamp	24.9286	**0.8102**	0.3319	24.0174	0.7875	0.3701	25.7934	0.8490	0.2998
Deformable-3DGS [50] w/ timestamp	23.4073	0.8058	**0.3305**	21.3836	0.7738	0.3801	24.1863	0.8427	**0.2865**

Method	Sword			Presentation			Average		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGStream [44]	24.0154	0.8284	0.2515	25.2535	0.8344	0.2866	22.7388	0.7630	0.3162
Deformable-3DGS [50]	21.2199	0.8603	0.2358	25.2457	0.8872	0.2372	23.0306	0.8329	0.2959
4DGaussian [49]	**25.4828**	**0.8668**	0.2671	28.2414	0.8994	0.2265	**25.7366**	**0.8432**	0.2983
4DGaussian [49] w/ timestamp	25.0993	0.8598	0.2778	**28.2761**	**0.8995**	**0.2262**	25.6230	0.8412	0.3012
Deformable-3DGS [50] w/ timestamp	20.9872	0.8612	**0.2345**	24.5876	0.8857	0.2370	22.9104	0.8338	**0.2937**

3.3 Data Processing

Video Post-processing. To ensure visual consistency across multi-view egocentric videos, we applied Adobe Premiere Pro [38] post-processing to standardize visual characteristics across all scenes. This included white balance calibration, exposure adjustment, and flicker removal from artificial lighting to eliminate sensor-specific color biases. Global exposure compensation and contrast adjustments further improved brightness uniformity while retaining shadow details.

The pose estimation process consists of two components: monocular pose tracking and multi-camera pose synthesis.

Monocular Pose Tracking. We applied a comprehensive set of state-of-the-art (SOTA) algorithms and engineering-validated classical methods to each camera's footage within every scene. Following the experimental benchmarks from Camerabench [30], we selected representative approaches including Anycam [48], MegaSAM [29], CUT3R [47], MonST3R [54], PySLAM [12], covering both dynamic scene reconstruction (SfM) and simultaneous localization and mapping (SLAM) frameworks. We performed spherical linear interpolation [19] on the sensor's rotational quaternion data to estimate the rotational pose at each video frame capture moment. We perform data fusion [1] aligning all estimated trajectories with the interpolated sensor data. This alignment process utilized the 3-DoF rotational pose data from the official SDK to compute the relative 6-DoF camera poses for each view.

Multi-camera Pose Synthesis. We instructed all performers to fixate on a common object in the first frame, establishing a static reference scene. For this initial static scene in the first frame, we captured extensive supplementary images from various angles to enrich the scene details and ensure robust reconstruction. All images were processed through the structure-from-motion pipeline in COLMAP [41] to obtain initial camera poses for all views. To maintain scale consistency across traces in monocular tracking, we selected an additional keyframe per view that contained content similar to the first frame but with noticeable translation. The displacement between these paired frames provided scale constraints to normalize translation components across all camera views. After completing all the preparation steps, we integrated the poses of all frames within the same scene into absolute poses for scene reconstruction through global motion-guided pose synthesis. Specifically, we applied the following procedure: 1) Accumulated relative poses were anchored to the initial poses in COLMAP-reconstructed of each view, 2) Translation components were scaled using the displacement ratios from paired keyframes. Experimental validation demonstrated that the resulting camera poses achieved sufficient accuracy for scene reconstruction tasks. Our data acquisition and processing pipeline is shown in Figure 3.

4 Experiment

4.1 Baselines

Given the absence of prior literature on multi-egocentric dynamic scene reconstruction at the time of our dataset release, we therefore focus on adapting methods originally developed for general dynamic scene reconstruction, including deformation-filed-based methods 4DGaussian [49] and Deformable-3DGS (D-3DGS) [50], and a streaming method 3DGSteam [44].

Due to the lack of existing datasets, these baselines cannot be directly applied to our multi-egocentric dynamic scene reconstruction dataset. To evaluate our dataset's effectiveness with these baselines, we modified their data loading pipelines through a hard-coded approach that directly accesses per-frame images and camera poses. This modification preserves the core reconstruction mechanisms of the baselines while revealing their genuine performance in multi-egocentric dynamic scene reconstruction scenarios.

4.2 Implementation Settings

We adopt PSNR, SSIM, and LPIPS as quantitative evaluation metrics, which are widely used in scene reconstruction research. For all baselines, we adopt their officially recommended default settings. These methods generally initialize scene representations using observations from the first frame and predict motions of Gaussian primes, which works effectively for datasets with fixed or small-range camera movements. However, our dataset features large-range camera rotations that result in scene extents significantly exceeding the

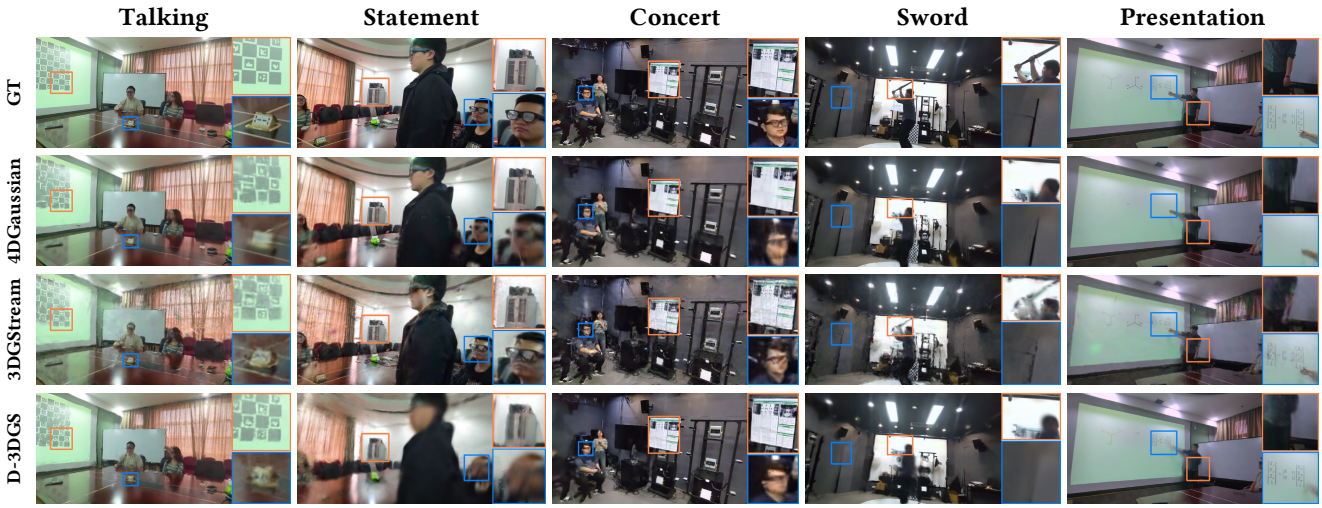

Figure 4: The visualization results from our validation experiments demonstrate that baselines employing different reconstruction strategies exhibit distinct characteristics.

visible range captured in any single frame. When extensive new scene regions emerge across viewpoint changes, this can severely impact performance of the method with the initialization from a single frame. To mitigate this effect, we applied random point cloud initialization within the complete scene space for all baselines. In particular, we employ all available views to perform 3DGS [22] pipeline on the randomly generated point cloud, thereby obtaining the full scene's Gaussian representation serving as the initialization for 3DGStream. Employing better initialization strategies tailored to our dataset may yield improved reconstruction outcomes.

4.3 Experiment Results

Quantitative Results. The quantitative experimental results of all selected baselines are summarized in Table 2. The result shows that the ***sword*** and ***presentation*** scene exhibit relatively smaller camera rotational motion magnitudes and more static background appearances, resulting in minimal novel background regions requiring reconstruction, which hence demonstrates relatively favorable quantitative results across methods. In contrast, the ***statement*** scene presents large-scale camera rotations/translations combined with extensive specular reflections, making it the most technically challenging scenario among our dataset.

Visualization Results. The visualization results of our validation experiments are presented in Figure 4, where 4DGaussian demonstrates superior static scene reconstruction capabilities that align with its favorable quantitative metrics, but exhibits limitations in capturing high dynamic object. This performance discrepancy may stem from the MLP-based deformation field prediction mechanism in 4DGaussian, which prioritizes low-frequency information processing and consequently generates overly smoothed reconstructions. In contrast, 3DGStream produces reconstructions containing high-frequency noise but achieves the best preservation of high-dynamic details among all baselines, which is evident in sword prop reconstruction in the ***sword*** scene. For multi-egocentric dynamic scene reconstruction, both high-quality background and accurate dynamic object are critical requirements. The trade-off between these two aspects constitutes a critical research challenge that must be addressed in future studies aiming at this task.

Experiment About Timestamp. As described in Section 3.3, we captured timestamps per frame. The collected timestamps closely align with the theoretical 30 FPS video capture intervals. Nevertheless, we conducted experiments incorporating actual timestamps as view-specific input in 4DGaussian and Deformable-3DGS, with results shown in "4DGaussian w/ timestamp" and "Deformable-3DGS w/ timestamp" of Table 2. Experimental results demonstrate that timestamps exert varying impacts across different scenes and methods, with performance improvements observed in some cases and degradations in others. Incorporating timestamps transforms the optimization from a single-scene estimation across five views to individual scene estimations per view with more dense time sampling. Although such distinctions exist, quantitative experiments demonstrate that this discrepancy is statistically insignificant.

5 Conclusion

We present MultiEgo, the first dynamic scene reconstruction dataset composed of multi-egocentric perspectives. Compared with previous dynamic scene reconstruction datasets and egocentric video collections, MultiEgo provides essential data elements for dynamic scene reconstruction tasks, including accurate camera pose annotations and synchronized temporal alignment. The dataset incorporates various challenges commonly encountered in multi-egocentric dynamic reconstruction scenarios, such as high dynamic objects and complex lighting conditions. We provide comprehensive descriptions of the dataset's contents. In validation experiments, we demonstrate the effectiveness of the data set through several baselines. In addition, we conducted additional studies to evaluate the impact of directly captured frame timestamps on reconstruction performance.

Acknowledgments

This work was partly supported by the NSFC62431015, Science and Technology Commission of Shanghai Municipality No.24511106200, the Fundamental Research Funds for the Central Universities, Shanghai Key Laboratory of Digital Media Processing and Transmission under Grant 22DZ2229005, 111 project BP0719010.

References

[1] Athira Chandra Babu, Ravi Kumar Karri, et al. 2018. Sensor data fusion using Kalman filter. In 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C). IEEE, 29–36.

[2] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. 2015. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. In The IEEE International Conference on Computer Vision (ICCV).

[3] Aayush Bansal and Michael Zollhoefer. 2023. Neural pixel composition for 3d-4d view synthesis from multi-views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 290–299.

[4] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive light field video with a layered mesh representation. ACM Trans. Graph. 39, 4, Article 86 (Aug. 2020), 15 pages. doi:10.1145/3386569.3392485

[5] Ang Cao and Justin Johnson. 2023. HexPlane: A Fast Representation for Dynamic Scenes. CVPR (2023).

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2021. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. International Journal of Computer Vision (IJCV) (2021). https://doi.org/10.1007/s11263-021-01531-2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In European Conference on Computer Vision (ECCV).

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2021. The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 43, 11 (2021), 4125–4141. doi:10.1109/TPAMI.2020.2991965

[9] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. 2009. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. (2009).

[10] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. 2022. Fast dynamic radiance fields with time-aware neural voxels. In SIGGRAPH Asia 2022 Conference Papers. 1–9.

[11] Alircza Fathi. 2012. Social interactions: A first-person perspective. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '12). IEEE Computer Society, USA, 1226–1233.

[12] Luigi Freda. 2025. pySLAM: An open-source, modular, and extensible framework for SLAM. arXiv preprint arXiv:2502.11955 (2025).

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In IEEE/CVF Computer Vision and Pattern Recognition (CVPR).

[14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 19383–19400.

[15] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. 2025. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. arXiv preprint arXiv:2505.11709 (2025).

[16] Qiang Hu, Qihan He, Houqiang Zhong, Guo Lu, Xiaoyun Zhang, Guangtao Zhai, and Yanfeng Wang. 2025. VARFVV: View-Adaptive Real-Time Interactive Free-View Video Streaming With Edge Computing. IEEE Journal on Selected Areas in Communications 43, 7 (2025), 2620–2634. doi:10.1109/JSAC.2025.3559140

[17] Qiang Hu, Zihan Zheng, Houqiang Zhong, Sihua Fu, Li Song, Guangtao Zhai, Yanfeng Wang, et al. 2025. 4DGC: Rate-Aware 4D Gaussian Compression for Efficient Streamable Free-Viewpoint Video. arXiv preprint arXiv:2503.18421 (2025).

[18] Qiang Hu, Houqiang Zhong, Zihan Zheng, Xiaoyun Zhang, Zhengxue Cheng, Li Song, Guangtao Zhai, and Yanfeng Wang. 2025. VRVVC: Variable-Rate NeRF-Based Volumetric Video Compression. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39. 3563–3571.

[19] Mehdi Jafari and Habib Molaei. 2014. Spherical linear interpolation and Bézier curves. General Scientific Researches 2, 1 (2014), 13–17.

[20] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In The IEEE International Conference on Computer Vision (ICCV).

[21] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2017. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[23] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. 1346–1353. doi:10.1109/CVPR.2012.6247820

[24] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. 2022. Streaming radiance fields for 3d video synthesis. Advances in Neural Information Processing Systems 35 (2022), 13485–13498.

[25] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. 2022. Neural 3D Video Synthesis From Multi-View Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5521–5531.

[26] Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In Proceedings of the European conference on computer vision (ECCV). 619–635.

[27] Yin Li, Zhefan Ye, and James M Rehg. 2015. Delving into egocentric actions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 287–295.

[28] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2024. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8508–8520.

[29] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. 2024. MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos. arXiv preprint (2024).

[30] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. 2025. Towards Understanding Camera Motions in Any Video. (2025).

[31] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2022. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 21013–21022.

[32] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2013), 2714–2721. doi:10.1109/CVPR.2013.350 Copyright: Copyright 2013 Elsevier B.V., All rights reserved.; 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013 ; Conference date: 23-06-2013 Through 28-06-2013.

[33] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In 3DV.

[34] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. 2020. You2me: Inferring body pose in egocentric video via first and second person interactions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 9890–9900.

[35] Curtis G Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. 2020. Egocom: A multi-person multi-modal egocentric communications dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 6 (2020), 6783–6793.

[36] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. ACM Trans. Graph. 40, 6, Article 238 (dec 2021).

[37] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2847–2854. doi:10.1109/CVPR.2012.6248010

[38] Adobe Premiere Pro. 2018. Adobe Premiere Pro.

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2020. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[40] Neus Sabater, Guillaume Boisson, Benoit Vandame, Paul Kerbiriou, Frederic Babon, Matthieu Hog, Remy Gendrot, Tristan Langlois, Olivier Bureller, Arno Schubert, et al. 2017. Dataset and pipeline for multi-view light-field video. In Proceedings of the IEEE conference on computer vision and pattern recognition Workshops. 30–40.

[41] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In Conference on Computer Vision and Pattern Recognition (CVPR).

[42] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018).

[43] Tomas Simon, Hanbyul Joo, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. CVPR (2017).

[44] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 2024. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 20675–20685.

[45] Yansong Tang, Zian Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2019. Multi-Stream Deep Neural Networks for RGB-D Egocentric Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology 29, 10 (2019), 3001–3015. doi:10.1109/TCSVT.2018.2875441

[46] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. 2023. Ego-Only: Egocentric Action Detection without Exocentric Transferring. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 5250–5261.

[47] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. 2025. Continuous 3D Perception Model with Persistent State. In CVPR.

[48] Felix Wimbauer, Weirong Chen, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. 2025. AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[49] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 20310–20320.

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

[51] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. 2024. Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting. International Conference on Learning Representations (ICLR).

[52] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. 2020. Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera. (June 2020).

[53] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. 2024. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. arXiv preprint arXiv:2406.19811 (2024).

[54] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. 2024. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. arXiv preprint arxiv:2410.03825 (2024).

[55] Zihan Zheng, Houqiang Zhong, Qiang Hu, Xiaoyun Zhang, Li Song, Ya Zhang, and Yanfeng Wang. 2024. HPC: Hierarchical progressive coding framework for volumetric video. In Proceedings of the 32nd ACM International Conference on Multimedia. 7937–7946.

[56] Zihan Zheng, Houqiang Zhong, Qiang Hu, Xiaoyun Zhang, Li Song, Ya Zhang, and Yanfeng Wang. 2024. JOINTRF: End-To-End Joint Optimization for Dynamic Neural Radiance Field Representation and Compression. In 2024 IEEE International Conference on Image Processing (ICIP). 3292–3298. doi:10.1109/ICIP51287.2024.10647336