# `Aria-NeRF`: Multimodal Egocentric View Synthesis

*Abstract*— **We seek to accelerate research in developing rich, multimodal scene models trained from egocentric data, based on differentiable volumetric ray-tracing inspired by Neural Radiance Fields (NeRFs). The construction of a NeRF-like model from an egocentric image sequence plays a pivotal role in understanding human behavior and holds diverse applications within the realms of VR/AR. Such egocentric NeRF-like models may be used as realistic simulations, contributing significantly to the advancement of intelligent agents capable of executing tasks in the real-world. The future of egocentric view synthesis may lead to novel environment representations going beyond today's NeRFs by augmenting visual data with multimodal sensors such as IMU for ego-motion tracking, audio sensors to capture the surface texture and human language context, and eye-gaze trackers to infer human attention patterns in the scene. To support and facilitate the development and evaluation of egocentric multimodal scene modeling, we present a comprehensive multimodal egocentric video dataset. This dataset offers a comprehensive collection of sensory data, featuring RGB images, eye-tracking camera footage, audio recordings from a microphone, atmospheric pressure readings from a barometer, positional coordinates from GPS, connectivity details from Wi-Fi and Bluetooth, and information from dual-frequency IMU datasets (1kHz and 800Hz) paired with a magnetometer. The dataset was collected with the Meta Aria Glasses wearable device platform. We evaluated two baseline NeRF-based models, Nerfacto and NeuralDiff, on our dataset. While they were capable of producing reasonable visual reproduction of the scene, our findings also highlight opportunities for further improvement using a variety of sensing modalities beyond vision. The diverse data modalities and the real-world context captured within this dataset serve as a robust foundation for furthering our understanding of human behavior and enabling more immersive and intelligent experiences in the realms of VR, AR, and robotics.**

## I. Introduction

Recent advances in VR/AR technologies highlight a growing need for creating immersive virtual environments. Neural Radiance Fields (NeRFs) [21] is a technique that has gained much attention for its capability to generate photorealistic 3D scenes, meeting this demand for strengthened immersion and realism. Utilizing NeRF for the creation of lifelike simulations is of significant value, particularly in the development of intelligent agents capable of executing real-world tasks. Nevertheless, dynamic NeRF remains a complex problem [15]. Unlike traditional NeRF, which deals with static scenes, dynamic NeRF aims to capture and represent objects and scenes that change over time. This introduces the need for modeling complex temporal dynamics, such as object motion, deformation, or interactions, which can be challenging to represent accurately. On the other hand, it also presents an intriguing avenue to explore whether multimodal sensory data can enhance NeRF training. This work focuses on egocentric view synthesis, which is a natural scenario featuring rich multimodal data that can be captured by multi-sensory wearable devices.

The progression of egocentric vision [28, 31] heavily relies on hardware advancements. This is particularly relevant in the context of wearable devices like Aria Glasses [24, 33], which capture data from a first-person perspective in real-life scenarios. Aria Glasses, designed as a research tool to accelerate advances in AR/VR, embodied AI, and human behavior modeling, employ a range of sensors to capture first-person perspective video, audio, data on eye movement and location, providing a comprehensive platform for understanding a user's intention, as well as their interactions with the world.

In this paper, we present the `Aria-NeRF Dataset`, a comprehensive multimodal egocentric dataset designed for multimodal egocentric scene modeling, with diverse real-world multi-sensory data captured using Aria Glasses. Extensive experiments demonstrate that our dataset is a rich testbed for typical NeRF-based tasks and algorithms, with high-quality annotations to explore by future research. Our dataset is also scalable, offering a cost-effective pipeline for converting Aria Glasses-captured videos into NeRF-compatible training data.

In summary, our main contributions are as follows:

- We introduce the task of `Multimodal Egocentric Scene Modeling`, a step towards egocentric view synthesis with neural scene representations, using Fisheye RGB images and multimodal sensory data.
- We build a novel `Aria-NeRF Dataset`, which includes multiple modalities, such as Fisheye images, RGB, depth, IMU, audio, etc. The proposed `Aria-NeRF Dataset` serves as a rich testbed for advancing multimodal NeRF, for example, audio-guided NeRF, and gaze-guided NeRF. In particular, `Aria-NeRF Dataset` has language annotations, which are suitable for training LLMs-guided NeRF.
- We evaluate and benchmark Nerfacto [35] and NeuralDiff [37] on the proposed `Aria-NeRF Dataset` extensively. The results reveal the challenging nature of our dataset and egocentric view synthesis, suggesting the need for further improvement of the current NeRF methods.

## II. Related Works

*a) One/Few-shot NeRF:* The field of one/few-shot Neural Radiance Fields (NeRF) has seen significant advancements in recent years. Several approaches have been proposed to tackle the challenges of synthesizing novel views and reconstructing 3D scenes with limited available data.

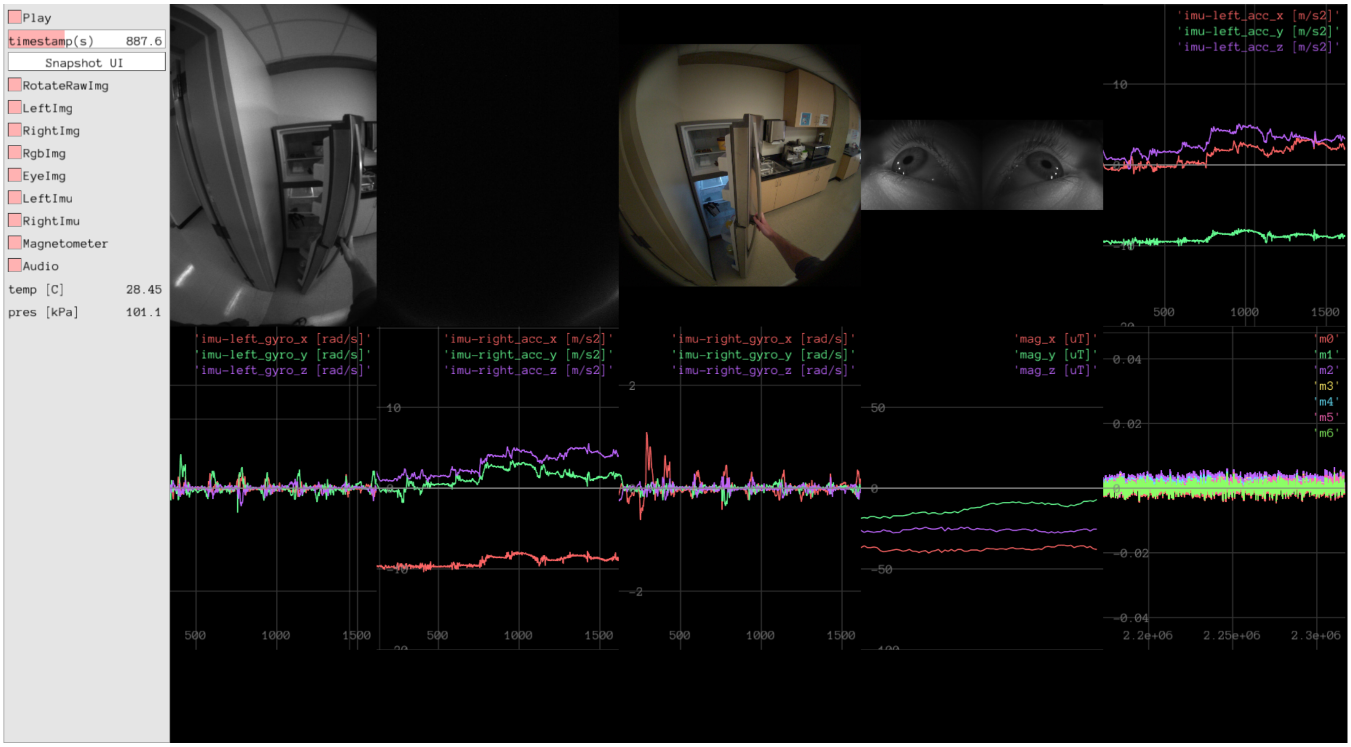

Fig. 1: `Aria-NeRF Dataset`, Kitchen 1 subset, comprises a diverse range of sensory data, including RGB images, ET camera, microphone, barometer, GPS, Wi-Fi, Bluetooth, SLAM, and two sets of IMU data (1kHz and 800Hz), along with a magnetometer.

Mip-NeRF 360 [2] addresses the task of Unbounded Anti-Aliased Neural Radiance Field synthesis, using a non-linear scene parameterization, online distillation, and a distortion-based regularizer to overcome the challenges presented by unbounded scenes. EgoNeRF [5] employs spherical coordinates and leverages 360-degree panoramic videos as input to construct neural radiance fields. In addition, there have been several multi-stage approaches [3, 12] that attempt to synthesize new view images by reconstructing an explicit mesh from egocentric omnidirectional videos. Zero-1-to-3 [17] enables zero-shot novel view synthesis and 3D reconstruction using only a single image.

In contrast, we assume a more casual input setting, where the viewpoint and scene composition may vary widely, data may be abundant, but from an egocentric viewpoint embedded in the scene, and video data is augmented by other multimodal data sources.

*b) Dynamic NeRF:* Recent studies have also focused on synthesizing novel views of dynamic scenes using a single camera. D-NeRF [27] has the capability to synthesize novel views of dynamic scenes with intricate non-rigid geometries at arbitrary time points. Nerfies [26] and HyperNeRF [25] represent scenes using deformation fields that are conditioned on either time instant [27] or per-frame learned latent deformation cod [25, 26, 36]. NeuralDiff [37] tackles 3D object segmentation by employing a triple-stream neural renderer to separate the background, foreground, and actor. While these methods can handle lengthy videos, their primary effectiveness lies in object-centric scenes with limited object motion and controlled camera paths. Alternatively, some approaches model scenes as time-varying NeRFs [8, 9, 14, 41, 43]. NSFF [14] employs neural scene flow fields to capture complex 3D scene motion in real-world videos. However, it performs best on short, forward-facing videos lasting 1-2 seconds duration. DynIBaR [15], focuses on synthesizing novel views from monocular videos depicting complex dynamic scenes. Our dataset is also well-suited for dynamic NeRF tasks.

*c) Multimodal NeRF:* Multimodal Neural Radiance Field [44] is valuable for robot vision and scene understanding. Zhu et al. [48] introduce a method that aligns different modalities, incorporating point clouds and infrared image supervision. CLIP-NeRF [40] proposes a unified framework that enables user-friendly manipulation of NeRF using either a short text prompt or an exemplar image. MMNeRF [45] learns multimodal and multi-view features to guide neural radiance fields toward a generic model. OMMO [19] serves as a multimodal benchmark for outdoor NeRF-based tasks, providing complex objects and scenes with calibrated images, point clouds, and prompt annotations. ObjectFolder [10] is a dataset designed for multisensory object-centric learning, incorporating vision, audio, and touch modalities. In addition to RGB cameras, our dataset includes a range of sensor types such as Fisheye cameras, IMU (Inertial Measurement Unit), audio, and more. This diverse collection of data enables the training of NeRF models using multiple modalities.

*d) VR/AR:* NeRF holds great potential for creating immersive environments in Augmented and Virtual Reality

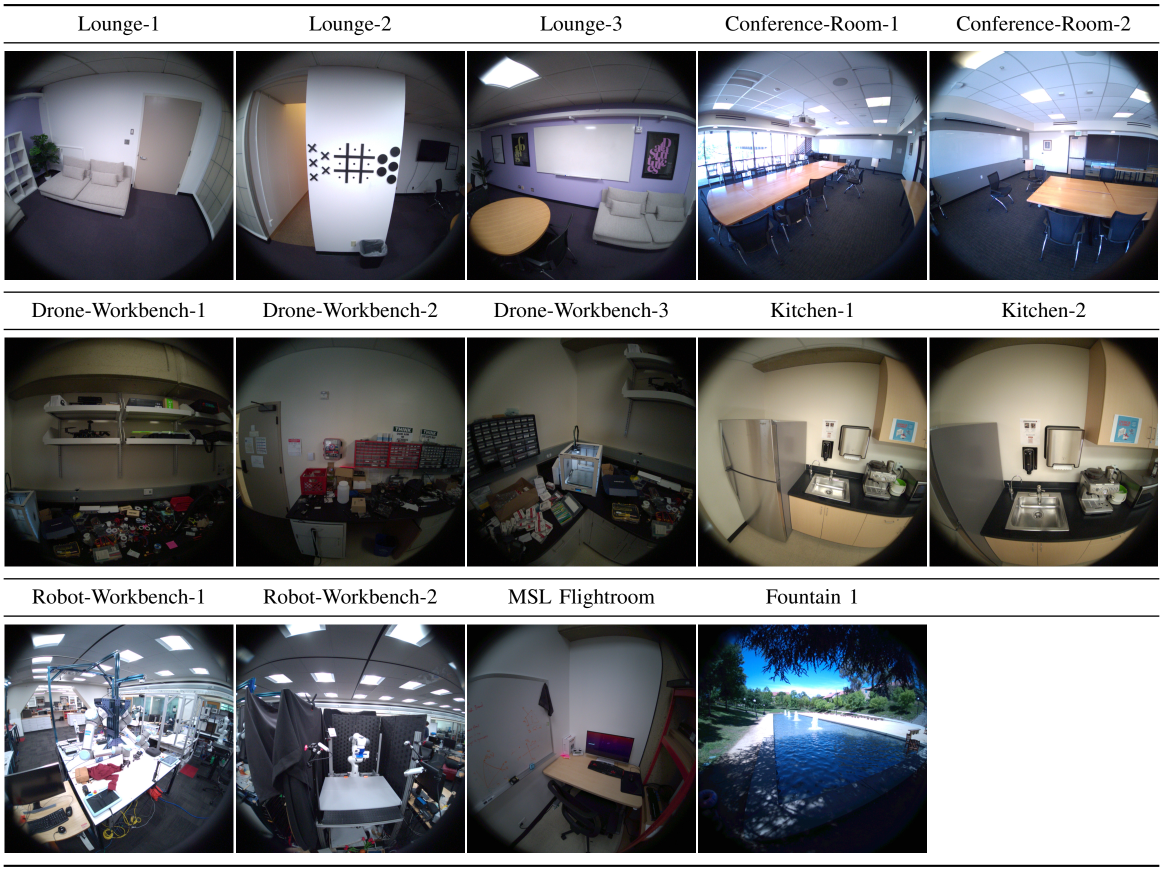

Fig. 2: Scene Examples

TABLE I: Dataset Statistics 1: Time and Sensors - RGB, Eye Tracking (ET), Microphone, Barometer, and GPS Data.

| Subset | Time (s) | RGB | ET | Microphone | Barometer | GPS |
|---|---|---|---|---|---|---|
| Fountain1 | 124.9 | 1261 | 1261 | 2953 | 6276 | 127 |
| Lounge-1 | 69.0 | 701 | 701 | 1640 | 3497 | 0 |
| Lounge-2 | 69.6 | 707 | 707 | 1656 | 3524 | 0 |
| Lounge-3 | 146.5 | 1476 | 1476 | 3457 | 7374 | 0 |
| Flightroom | 159.6 | 1607 | 1607 | 3765 | 8046 | 0 |
| Drone-Workbench-1 | 145.3 | 1465 | 1465 | 3430 | 7316 | 0 |
| Drone-Workbench-2 | 127.5 | 1286 | 1287 | 3013 | 6411 | 0 |
| Drone-Workbench-3 | 78.2 | 793 | 793 | 1856 | 3960 | 0 |
| Conference-Room-1 | 107.4 | 1085 | 1085 | 2541 | 5412 | 0 |
| Conference-Room-2 | 77.2 | 783 | 783 | 1833 | 3907 | 0 |
| Kitchen-1 | 134.8 | 1360 | 1360 | 3185 | 6791 | 0 |
| Kitchen-2 | 123.4 | 1245 | 1245 | 2917 | 6216 | 0 |
| Robot-Workstation-1 | 69.7 | 708 | 708 | 1657 | 3530 | 71 |
| Robot-Workstation-2 | 41.5 | 426 | 426 | 996 | 2122 | 43 |

(AR/VR) applications. NeRF is highly applicable with the ability to generate realistic and high-quality visual experiences in these domains. Fov-NeRF [7] is a technique that specifically targets Virtual Reality (VR) applications by introducing a gaze-contingent neural radiance field. This method enhances the responsiveness of neural synthesis within the VR environment, resulting in improved visual quality and realism in virtual experiences. Instant-3D [13] is an algorithm-hardware co-design acceleration framework that enables instant on-device NeRF training. This framework facilitates instant 3D reconstruction for AR/VR applications, allowing for real-time and interactive experiences. TLIO [18], trained with pedestrian data from a headset, has the capability to produce statistically consistent measurements and uncertainty for IMU-only state estimation. This contributes to accurate tracking and positioning in AR/VR scenarios. EPIC Fields [38] enhances the EPIC-KITCHENS dataset by incorporating 3D camera information. Recently, HoloAssist [42] is an egocentric human interaction dataset, where two people collaboratively complete physical manipulation tasks. Our dataset, which includes a commodity omnidirectional camera with two fish-eye lenses, has the potential to enhance VR/AR applications by providing valuable data for training and improving NeRF-based techniques.

TABLE II: Dataset Statistics 2: encompasses a diverse array of sensory data including Wi-Fi, Bluetooth, SLAM, and measurements from both IMUs and magnetometers. The dataset features two variants of IMU data: one sampled at 1kHz and another at 800Hz.

| Subset | Wi-Fi | Bluetooth | SLAM | IMU (1kHz) | IMU (800Hz) | Magnetometer |
|---|---|---|---|---|---|---|
| Fountain1 | 340 | 1 | 1261 | 126029 | 102124 | 1262 |
| Lounge-1 | 326 | 26 | 700 | 70037 | 56751 | 702 |
| Lounge-2 | 294 | 37 | 707 | 70683 | 57275 | 708 |
| Lounge-3 | 583 | 81 | 1476 | 147492 | 119508 | 1480 |
| Flightroom | 476 | 0 | 1607 | 160611 | 130143 | 1613 |
| Drone-Workbench-1 | 530 | 0 | 1465 | 146314 | 118556 | 1468 |
| Drone-Workbench-2 | 408 | 0 | 1286 | 128586 | 104195 | 1288 |
| Drone-Workbench-3 | 250 | 0 | 793 | 79241 | 64211 | 794 |
| Conference-Room-1 | 391 | 34 | 1085 | 108440 | 87871 | 1086 |
| Conference-Room-2 | 157 | 20 | 783 | 78257 | 63410 | 784 |
| Kitchen-1 | 522 | 0 | 1360 | 135900 | 110119 | 1363 |
| Kitchen-2 | 606 | 0 | 1245 | 124453 | 100845 | 1248 |
| Robot-Workstation-1 | 543 | 0 | 708 | 70760 | 57338 | 708 |
| Robot-Workstation-2 | 318 | 0 | 426 | 42562 | 34490 | 426 |

TABLE III: Quantitative Results. In the context of PSNR, SSIM, and LPIPS metrics, NeuralDiff generally surpasses Nerfacto across various scenarios.

| Subset | Nerfacto | | | NeuralDiff | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM ↑ | LPIPS ↓ | PSNR↑ | SSIM ↑ | LPIPS ↓ |
| Fountain1 | 20.16 | 0.7075 | 0.5212 | 29.09 | 0.9131 | 0.1695 |
| Lounge-1 | 19.93 | 0.7284 | 0.5261 | 29.98 | 0.9270 | 0.2238 |
| Lounge-3 | 19.63 | 0.7036 | 0.5951 | 29.93 | 0.9230 | 0.1441 |
| Flightroom | 19.41 | 0.7177 | 0.4601 | 29.30 | 0.8979 | 0.1689 |
| Drone-Workbench-1 | 20.52 | 0.6820 | 0.5626 | 25.67 | 0.8578 | 0.3315 |
| Drone-Workbench-2 | 20.93 | 0.6887 | 0.5169 | 30.60 | 0.9324 | 0.1149 |
| Drone-Workbench-3 | 25.05 | 0.7736 | 0.4804 | 31.28 | 0.9427 | 0.1050 |
| Conference-Room-1 | 17.10 | 0.6247 | 0.6767 | 19.22$^{\S}$ | 0.6140$^{\S}$ | 0.6908$^{\S}$ |
| Conference-Room-2 | 20.70 | 0.7267 | 0.5482 | 29.42 | 0.9444 | 0.1071 |
| Kitchen-1 | 20.19 | 0.6957 | 0.5854 | 32.72 | 0.9454 | 0.1144 |
| Kitchen-2 | 22.67 | 0.7538 | 0.4825 | 34.13 | 0.9634 | 0.0552 |
| Robot-Workstation-1 | 20.37 | 0.7336 | 0.4149 | 22.90 | 0.8474 | 0.2020 |
| Robot-Workstation-2 | 20.97 | 0.7420 | 0.4493 | 20.44 | 0.7779 | 0.3362 |

$^{\S}$ stops at epoch 4 due to convergence issues.

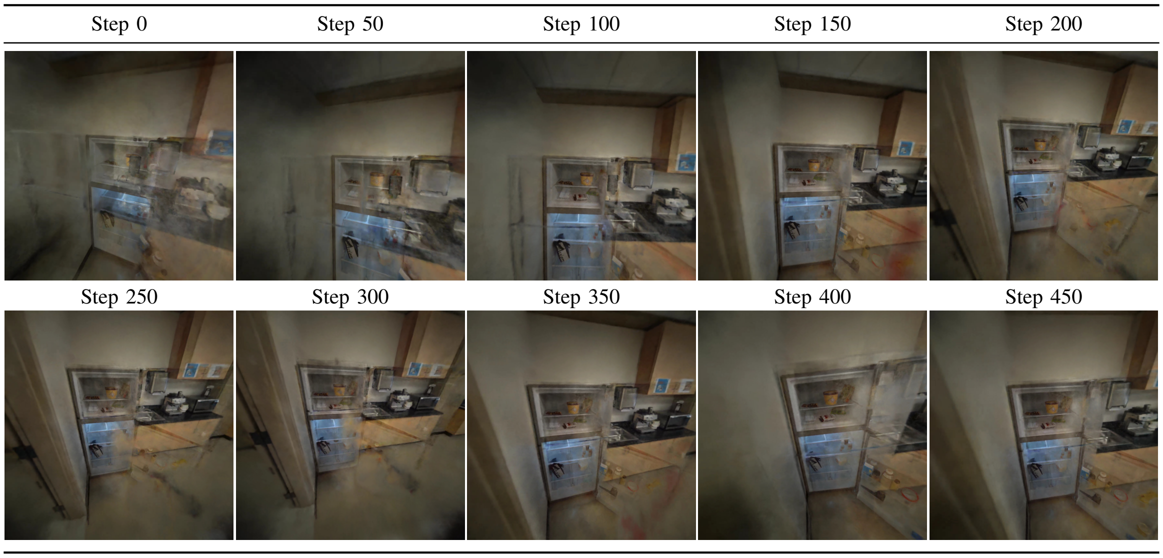

Fig. 3: Nerfacto Visualization Results on Kitchen 1 subset. We show the step numbers in the rendered video. Nerfacto results reveal some blurred regions, underscoring inherent limitations in its performance.

## III. Method

We evaluate two existing methods: Nerfacto [35] which is adept at constructing neural radiance fields for static scenes

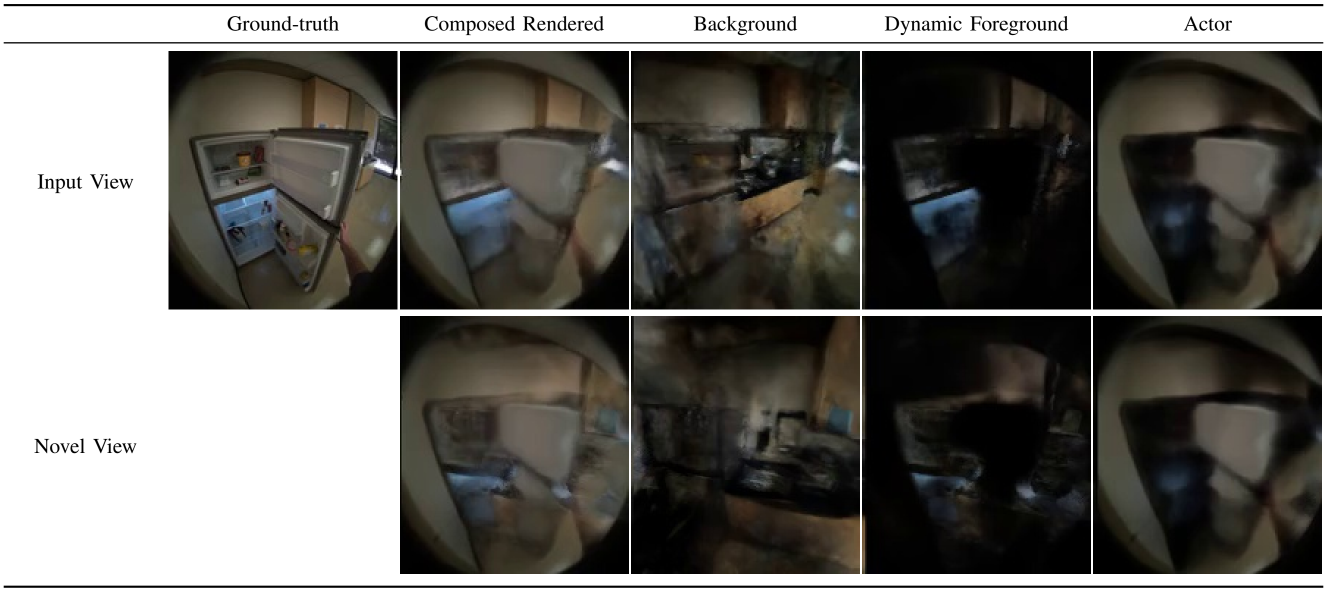

Fig. 4: NeuralDiff Visualization Results on Kitchen 1 subset. NeuralDiff can disentangle the background, dynamic foreground, and actors, all achieved in an unsupervised manner.

using real-world data, and NeuralDiff [37], a method explicitly tailored for dynamic Neural Radiance Field (NeRF) scenarios.

### A. Nerfacto

The Nerfacto model is the default model used by nerfstudio [35] for building neural radiance fields of static scenes from real data. This model is a combination of various established methods known for their efficacy with real data. Key techniques integrated into the Nerfacto model include camera pose refinement, per-image appearance conditioning, proposal sampling, scene contraction, and hash encoding.

*a) Ray Generation and Sampling:* The Nerfacto algorithm begins by optimizing camera views via an optimized SE(3) transformation [16]. Utilizing these views, RayBundles are generated. To enhance both the efficiency and efficacy of the sampling process, a piece-wise sampler is employed. Initially, this sampler operates uniformly up to a designated distance from the camera. Subsequently, its sampling becomes progressively distributed, with each sample's step size increasing incrementally. These samples are introduced to a proposal network sampler, as conceptualized in the MipNeRF-360 approach [1]. Nerfacto incorporates a compact fused MLP equipped with hash encoding [23] for representing the scene's density function, attributing to its computational efficiency without compromising accuracy. To further reduce the number of samples along rays, the proposal network sampler is designed to encompass multiple density fields.

*b) Scene Contraction and NeRF Field:* Many real-world scenes are unbounded, meaning they could extend indefinitely. Nerfacto applies scene contraction to transform this unbounded space into a fixed-size bounding box [1]. Instead of the conventional $L^2$ norm contraction, Nerfacto adopts an $L^\infty$ norm contraction, resulting in a cubic domain rather than a spherical boundary. This cubic form is more conducive to the voxel-based hash encodings. Subsequently, these compacted spatial samples are compatible with the hash encoding framework provided by Instant-NGP, accessible through the tiny-cuda-nn [22] Python interface. Additionally, Nerfacto integrates per-image appearance embeddings to mitigate variations in lighting and exposure encountered across different training cameras, referencing techniques used in [20]. It also incorporates strategies from Ref-NeRF [39] to enhance the computation and prediction of surface normals.

### B. NeuralDiff

NeuralDiff [37] is crafted for dynamic NeRF applications. It possesses the capability to autonomously disentangle the background, foreground, and actor within the NeRF representation.

NeuralDiff contains three sub-networks: Background density $\sigma_k^b \in \mathbb{R}_+$ and color $c_k^b \in \mathbb{R}^3$ can be obtained from the Background Network: $(\sigma_k^b, c_k^b) = \mathrm{MLP}^b(g_t r_k, d_t)$. The dynamic foreground is modelled by Foreground Network: $(\sigma_k^f, c_k^f, \beta_k^f) = \mathrm{MLP}^f(g_t r_k, z_t^f)$ produces a 'foreground' occupancy $\sigma^f$ and color $c^f$ . Additionally, it predicts an uncertainty score $\beta_k^f$. A frame-specific code $z^f \in \mathbb{R}^D$ captures the properties of the foreground that change over time. $z_t = B(t)\Gamma$ where $B(t) \in \mathbb{R}^P$ is a simple handcrafted basis and the motion $\Gamma \in \mathbb{R}^{P \times D}$ are coefficients such that $P \ll T$. Specifically, $B(t) = [1, t, \sin 2\pi t, \cos 2\pi t, \sin 4\pi t, \cos 4\pi t, \cdots]$ is a deterministic harmonic coding of time. Foreground objects are manipulated by the actor/observer, whose movements are sporadic, while the actor's body undergoes continuous motion. To model this dynamic actor, Actor Network $(\sigma_k^a, c_k^a, \beta_k^a) = \mathrm{MLP}^f(r_k, z_t^a)$ is used. The key difference is that the 3D point $r_k$ is expressed relative to the camera (v.s. $g_t r_k$ which is expressed relative to the world). $g_t \in SE(3)$ is the moving camera motion, where $SE(3)$ is the group of Euclidean transformations. $d_t$ is the unit-norm viewing direction.

### C. Dataset and Benchmark

Our dataset possesses three distinct characteristics:

- It comprises egocentric and dynamic scenes.
- The data is derived from real-world scenes.
- It incorporates multiple modalities.

The data modalities in our `Aria-NeRF Dataset` are captured by RGB cameras, ET camera, Microphone, Barometer, GPS, Wi-Fi, Bluetooth, SLAM/Mono Scene camera left, SLAM/Mono Scene camera right, IMU (1kHz), IMU (800Hz), and Magnetometer. The statistics for each subset and for the different modalities can be found in Tab. I and Tab. II.

*1) Data Collection:* Our dataset includes different scenarios. We collected multimodal sensory data in these scenarios, including RGB videos, ET camera, Microphone, Barometer, GPS, Wi-Fi, Bluetooth, SLAM, IMU (1kHz), IMU (800Hz), and Magnetometer, as shown in Figure 1. Each participant wears Aria Glasses to perform specific tasks within each scenario. These tasks may include activities like navigating, utilizing tools, and interacting with common household objects. In certain scenarios, GPS information is unavailable due to the absence of GPS signals within indoor environments. Similarly, the collection of Bluetooth data information is contingent upon the presence of a Bluetooth device near the scene. In the absence of a nearby Bluetooth device, capturing Bluetooth data becomes unfeasible.

In our data preprocessing pipeline, we employ Aria Data Tools[1] to extract data from MPS files[2], which is later utilized for visualization purposes. For pose estimation based on RGB video sequences, we employ COLMAP [32]. Our dataset does not necessitate additional annotations.

## IV. Experiments

In this section, we first introduce the training details. We then evaluate Nerfacto and NeuralDiff on our dataset. Our analysis reveals that the proposed `Aria-NeRF Dataset` is challenging, and there is much room for current NeRF methods to improve in the context of egocentric view synthesis.

### A. Baselines and Implementation Details

We run two baselines on our collected dataset:

- Nerfacto: We employed the default training settings of nerfstudio, conducting training for $30,000$ iterations with an initial learning rate of $10^{-8}$ during the prewarmup phase, and a final learning rate of $0.0001$.
- NeuralDiff: The NeuralDiff model is trained for 10 epochs with a learning rate of $0.0005$, utilizing 64 ray samples.

### B. Quantitative Results

Qualitative results are presented in Table III for Nerfacto and NeuralDiff. In terms of PSNR, SSIM, and LPIPS metrics, NeuralDiff generally surpasses Nerfacto across various scenarios. However, it is important to note that Nerfacto produces a de-distorted image, while NeuralDiff generates a fisheye image closely resembling the ground truth. The latter exhibits curvature characteristics.

### C. Qualitative Results

Qualitative results are presented in Figure 3 for Nerfacto and Figure 4 for NeuralDiff. Notably, the visualization of Nerfacto reveals some blurred regions, underscoring inherent limitations in its performance. In contrast, the visualization of NeuralDiff demonstrates its remarkable ability to disentangle elements within the scene, disentangling the background, dynamic foreground, and actors, all in an unsupervised manner. Both methods exhibit potential for enhancement in the application of dynamic NeRF.

## V. Discussion

An advantage of Aria Glasses, in contrast to the HoloLens, is their lightweight and highly portable design, making them seamlessly adaptable to people's everyday lives. It is worth noting that a limitation of the Aria Glasses RGB sensor is its relatively lower image resolution compared to current smartphone RGB cameras.

Recent advances in foundation models have brought new paradigm shifts and breakthroughs in different research areas [4, 29]. Foundation models emerge with generalist intelligence that can solve a wide range of tasks after being trained with a large quantity of data. With data at its core, foundation model research is embracing a new trend towards multimodality [30, 46]. `Aria-NeRF Dataset`, along with other large-scale egocentric datasets such as Ego4D [11] and EPIC-KITCHENS [6], holds great promise in bolstering the development of multimodal foundation models for egocentric view synthesis. Once trained, Aria-NeRF can also be used for many downstream perception and planning tasks, such as NeRF-based object detection [34], semantic segmentation [47], and so on.

## VI. Conclusion

In this work, we tackled the problem of egocentric view synthesis. To facilitate research in this field, we introduced `Aria-NeRF Dataset`, a multimodal egocentric dataset captured using Aria Glasses. We benchmarked two baseline models, Nerfacto and NeuralDiff, on this novel dataset. While these two models can generate reasonable view synthesis, the experimental results revealed much potential for improvement given the challenging nature of the dataset. The dataset's rich diversity of modalities and real-world context lay a solid groundwork for advancing our understanding of human behavior and bolstering more immersive and intelligent experiences in the realms of VR and AR.

[1] `https://facebookresearch.github.io/Aria_data_tools/`

[2] `https://facebookresearch.github.io/projectaria_tools/docs/data_utilities/core_code_snippets/mps`

## References

[1] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5470–5479.

[2] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *CVPR*, 2022.

[3] T. Bertel, M. Yuan, R. Lindroos, and C. Richardt, "OmniPhotos: Casual 360° VR photography," *ACM Transactions on Graphics*, vol. 39, no. 6, 266:1–12, Dec. 2020.

[4] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[5] C. Choi, S. M. Kim, and Y. M. Kim, "Balanced spherical grid for egocentric view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 590–16 599.

[6] D. Damen *et al.*, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 11, pp. 4125–4141, 2021.

[7] N. Deng *et al.*, "Fov-nerf: Foveated neural radiance fields for virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 11, pp. 3854–3864, 2022.

[8] C. Gao, A. Saraf, J. Kopf, and J.-B. Huang, "Dynamic view synthesis from dynamic monocular video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[9] H. Gao, R. Li, S. Tulsiani, B. Russell, and A. Kanazawa, "Monocular dynamic view synthesis: A reality check," in *NeurIPS*, 2022.

[10] R. Gao *et al.*, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *CVPR*, 2022.

[11] K. Grauman *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[12] H. Jang, A. Meuleman, D. Kang, D. Kim, C. Richardt, and M. H. Kim, "Egocentric scene reconstruction from an omnidirectional video," *ACM Trans. Graph.*, vol. 41, no. 4, Jul. 2022.

[13] S. Li *et al.*, "Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction," in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 2023, pp. 1–13.

[14] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6498–6508.

[15] Z. Li, Q. Wang, F. Cole, R. Tucker, and N. Snavely, "Dynibar: Neural dynamic image-based rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[17] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, *Zero-1-to-3: Zero-shot one image to 3d object*, 2023. arXiv: `2303.11328 [cs.CV]`.

[18] W. Liu *et al.*, "Tlio: Tight learned inertial odometry," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5653–5660, 2020.

[19] C. Lu, F. Yin, X. Chen, T. Chen, G. Yu, and J. Fan, "A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction," *arXiv preprint arXiv:2301.06782*, 2023.

[20] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *CVPR*, 2021.

[21] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[22] T. Müller, *tiny-cuda-nn*, version 1.7, Apr. 2021.

[23] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, 102:1–102:15, Jul. 2022.

[24] X. Pan *et al.*, "Aria digital twin: A new benchmark dataset for egocentric 3d machine perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 133–20 143.

[25] K. Park *et al.*, "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields," *ACM Trans. Graph.*, vol. 40, no. 6, Dec. 2021.

[26] K. Park *et al.*, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.

[27] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural Radiance Fields for Dynamic Scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[28] J. Qiu *et al.*, "Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8799–8806, 2022.

[29] J. Qiu *et al.*, "Large ai models in health informatics: Applications, challenges, and the future," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[30] J. Qiu *et al.*, "Visionfm: A multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence," *arXiv preprint arXiv:2310.04992*, 2023.

[31] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Naq: Leveraging narrations as queries to supervise episodic memory," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6694–6703.

[32] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] K. Somasundaram *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[34] J. Sun *et al.*, "Nerf-loc: Transformer-based object localization within neural radiance fields," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 5244–5250, 2023.

[35] M. Tancik *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.

[36] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 959–12 970.

[37] V. Tschernezki, D. Larlus, and A. Vedaldi, "Neuraldiff: Segmenting 3d objects that move in egocentric videos," in *2021 International Conference on 3D Vision (3DV)*, IEEE, 2021, pp. 910–919.

[38] V. Tschernezki *et al.*, "EPIC Fields: Marrying 3D Geometry and Video Understanding," 2023.

[39] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan, "Ref-NeRF: Structured view-dependent appearance for neural radiance fields," *CVPR*, 2022.

[40] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, "Clip-nerf: Text-and-image driven manipulation of neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3835–3844.

[41] L. Wang *et al.*, "Fourier plenoctrees for dynamic radiance field rendering in real-time," in *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 524–13 534.

[42] X. Wang *et al.*, "Holoassist: An egocentric human interaction dataset for interactive ai assistants in the real world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 20 270–20 281.

[43] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9421–9431.

[44] F. Zhan *et al.*, *Multimodal image synthesis and editing: The generative ai era*, 2023. arXiv: `2112.13592 [cs.CV]`.

[45] Q. Zhang, B. H. Wang, M. C. Yang, and H. Zou, "Mmnerf: Multi-modal and multi-view optimized cross-scene neural radiance fields," *IEEE Access*, vol. 11, pp. 27 401–27 413, 2023.

[46] Y. Zhang *et al.*, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023.

[47] S. Zhi, T. Laidlow, S. Leutenegger, and A. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[48] H. Zhu *et al.*, "Multimodal neural radiance field," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9393–9399.