Photoreal Scene Reconstruction from an Egocentric Device

ZHAOYANG LV, Reality Labs Research, Meta, USA
MAURIZIO MONGE, Reality Labs Research, Meta, United Kingdom
KA CHEN, Reality Labs Research, Meta, USA
YUFENG ZHU, Reality Labs Research, Meta, USA
MICHAEL GOESELE, Reality Labs Research, Meta, USA
JAKOB ENGEL, Reality Labs Research, Meta, USA
ZHAO DONG, Reality Labs Research, Meta, USA
RICHARD NEWCOMBE, Reality Labs Research, Meta, USA

Fig. 1. A comparison (from left to right) of (a) the raw camera held-out reference image, (b) the reconstructed view from a vanilla Gaussian-splatting algorithm using device calibration, (c) our reconstruction result using our proposed system, and (d) our reconstruction rendered with lifted dynamic range. We adjust the gamma of all comparisons for better visualization. Egocentric video may contain challenges in image quality due to high speed head motion and the form factor constraint. Our proposed system can recover photoreal scene reconstruction from the egocentric input videos with noises and heavy rolling-shutter effect. Note the improved clarity on the text in the low-light image areas due to our correct handling of rolling-shutter during both steps in visual inertial bundle adjustment and Gaussian-splatting. As a result, we can render the videos with higher dynamic range and further boost the details.

In this paper, we investigate the challenges associated with using egocentric devices to photorealistic reconstruct the scene in high dynamic range. Existing methodologies typically assume using frame-rate 6DoF pose estimated from the device's visual-inertial odometry system, which may neglect crucial details necessary for pixel-accurate reconstruction. This study presents two significant findings. Firstly, in contrast to mainstream work treating RGB camera as global shutter frame-rate camera, we emphasize the importance of employing visual-inertial bundle adjustment (VIBA) to calibrate the precise timestamps and movement of the rolling shutter RGB sensing camera in a high frequency trajectory format, which ensures an accurate calibration of the physical properties of the rolling-shutter camera. Secondly, we incorporate a physical image formation model based into Gaussian Splatting, which effectively addresses the sensor characteristics, including the rolling-shutter effect of RGB cameras and the dynamic ranges measured by sensors. Our proposed formulation is applicable to the widely-used variants of Gaussian Splats representation. We conduct a comprehensive evaluation of our pipeline using the open-source Project Aria device under diverse indoor and outdoor lighting conditions, and further validate it on a Meta Quest3 device. Across all experiments, we observe a consistent visual enhancement of +1 dB in PSNR by incorporating VIBA, with an additional +1 dB achieved through our proposed image formation model. Our complete implementation, evaluation datasets, and recording profile are available at https://www.projectaria.com/photoreal-reconstruction/

1 Introduction

Egocentric glasses equipped with first-person view cameras capture the world in an unparalleled perspective from the viewpoint of the human wearer. Scalable photorealistic 3D reconstruction from these inputs holds significant potential for various applications in augmented reality, robotics, and spatial artificial intelligence.

Seminal work in neural rendering [Kerbl et al. 2023; Mildenhall et al. 2021; Müller et al. 2022] has significantly advanced the field of photorealistic scene reconstruction using mobile cameras. Unlike traditional techniques that required depth sensing [Newcombe et al. 2011; Straub et al. 2019], neural rendering relies solely on posed images as input, making it well-suited to the form factor of egocentric glasses. Several studies have demonstrated that neural rendering facilitates 3D scene reconstruction and understanding [Gu et al. 2024; Tschernezki et al. 2023].

The constant motion of the human head presents significant challenges for image sensing and localization in egocentric devices, both of which can adversely affect the quality of neural reconstruction. Existing solutions employing visual-inertial odometry can provide a fast efficient solution for six degree of freedom (6DoF) tracking for the device, which is further used to estimate a frame-rate 6DoF pose for the RGB camera. However, a rolling-shutter camera is sensitive to the fast moving head and a frame-rate pose cannot accurately represent the correct pixel motions. Due to form factor constraints, the image quality captured by sensors lacking hardware stabilization can additionally be also compromised by potential motion blur and noise under undesired lighting conditions [Goesele et al. 2025].

In this paper, we examine the challenges associated with egocentric sensing and propose a systematic 3D reconstruction framework for photorealistic novel-view rendering. Our approach includes specific details on device calibration, reconstruction methods, and the capture process. We utilize Project Aria [Engel et al. 2023], an open-source egocentric glasses platform, for data capture, thereby providing representative data within appropriate form factor constraints. Furthermore, we demonstrate that insights gained from our study can be generalized to another commercial headset, such as the Meta Quest 3 device.

Our contributions are as follows:

Firstly, we address the importance of employing visual-inertial bundle adjustment (VIBA) that accounts for the rolling-shutter behavior of the RGB camera. This provides a continuous camera trajectory to model pixel movement in neural reconstruction. Our experiments demonstrate that using VIBA will consistently improve the novel view quality in Gaussian Splatting by +1db in PSNR.

Secondly, we introduce a rasterization-based image formation pipeline that addresses common artifacts in physical image formation, including rolling shutter, lens shading, exposure, and gain compensation. Our approach is distinct in that we represent image poses as posed pixel arrays sampled from a continuous trajectory, rather than assigning a single camera pose per image, and preserve the merit of Gaussian rasterization. Unlike existing methods that require ray-tracing Gaussians, e.g. [Moenne-Loccoz et al. 2024], our formulation is applicable to general-purpose rasterization-based Gaussian splatting. When being applied to 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023], our approach can further enhance reconstruction quality by +1 dB. We outperform existing baselines and demonstrate a large quality improvement in handling complex scenes observed by egocentric devices.

Third, to reduce the effect of blur with rapid head motion in darker indoor scenes, we propose a strategy of deliberately underexposing input videos during the capture, inspired by HDR+ [Hasinoff et al. 2016]. We demonstrate we can reconstruct high quality noise-free scene radiance from noisy dim input videos, and further render sharp blur-free videos at a higher dynamic range.

We evaluate our algorithm on recorded datasets across scenes with different scale and complexity. In addition to existing public datasets, we recorded a new Aria scene dataset following our capture process to benchmark this study. We release our code, dataset and capture profiles, with details at Project Aria Photoreal Reconstruction.

2 Related Work

Since the seminal work NeRF [Mildenhall et al. 2021] demonstrated novel view synthesis using radiance fields from posed images, numerous efforts have extended this to be faster [Fridovich-Keil et al. 2022; Müller et al. 2022], with anti-aliasing [Barron et al. 2022], at bigger scale [Barron et al. 2023] or dynamic scenes [Li et al. 2022]. Different from a raytracing formulation in NeRF, Gaussian Splatting [Kerbl et al. 2023] introduced a new rasterization based neural radiance fields composed of anisotropic 3D Gaussians (3D-GS) efficiently reconstructs scenes at high quality and its extension show superior performance for geometry reconstruction as well [Huang et al. 2024; Yu et al. 2024]. Recent methods [Condor et al. 2024; Mai et al. 2024; Moenne-Loccoz et al. 2024] use ray-tracing to optimize 3D-GS, addressing physical camera artifacts like rolling-shutter effects and deblurring, albeit with increased complexity. Our work addresses the physical camera properties using the rasterization based 3D-GS and shows it can improve the reconstruction quality on both 3D-GS and its variants.

Image blur and noise are common in mobile camera captures and have been extensively studied in classical [Hasinoff et al. 2016] and learning based methods [Chen et al. 2018]. Multi-view images aid in understanding image formation from its principle, with seminal work in Richardson-Lucy deblurring [Liu et al. 2010] and HDR+ [Hasinoff et al. 2016]. This concept extends to neural radiance fields. RawNeRF [Mildenhall et al. 2022] reconstructed radiance fields from

noisy low dynamic range input and outperformed single and multi-image raw denoisers. Approaches have explored reconstruction from blurred inputs using NeRF [Ma et al. 2021; Wang et al. 2023] or Gaussian splatting [Lee et al. 2024; Seiskari et al. 2024; Zhao et al. 2024], with a similar concept applicable to rolling-shutter as well [Niu et al. 2024]. They estimate the scene radiance with camera ray deformation induced by motion. However, due to unknown sensor calibration for exposure or rolling-shutter, they require jointly optimizing camera deformation along with the scene radiance. One recent approach [Seiskari et al. 2024] levarages high frequency visual inertial tracking and proposed to rasterize with additional pixel velocity measured against visual aligned IMU information.

Our approach is most related to the synthesized capabilities from RawNeRF [Mildenhall et al. 2022] and 3DGS-on-move [Seiskari et al. 2024]. Unlike RawNeRF, we used Gaussian Splatting for the scene representation and addressed the rolling-shutter in egocentric video beyond camera noise. Our solution is based on high-frequency tracking for the camera sensors. It addresses the pose acquisition issue to apply RawNeRF or its variants [Singh et al. 2024] in practice at scale. Our method leverages high-frequency tracking from VIO, explicitly modeling rolling-shutter and blur as in 3DGS-on-move [Seiskari et al. 2024]. Unlike 3DGS-on-move, which requires rasterizing pixel velocities with pinhole assumption for rolling-shutter and jointly pose optimization during training, our approach applies to any Gaussian Splatting variant with lens distortion, and we demonstrate superior quality consistently across scenes.

Neural rendering methods are commonly evaluated using static multi-view images captured by devices that uses image signal processing (ISP) [Barron et al. 2022, 2023; Mildenhall et al. 2019], which includes non-transparent processes of denoising, deblurring, sharpening and tone mapping that cannot be inverted. The camera calibration and poses are acquired using off-the-shelf structure from motion tools, such as COLMAP [Schönberger and Frahm 2016]. For images collected with extreme blur [Ma et al. 2021] or noise, the success in acquiring such camera ground truth can be extremely volatile. For successfully localized scenes, this calibration process also lacks physically meaningful calibration information to study the impact of 3D motions. 3DGS-on-the-move [Seiskari et al. 2024] collected hand-held videos with synchronized IMUs and demonstrated the first 3D-GS to jointly optimize rolling-shutter and motion-blur aware poses using imperfect factory calibration as initialization. Compared to all existing work, we propose to address these challenges in an egocentric device systematically from the input capture procedure to reconstruction algorithm. We evaluate the system robustly across different scenarios and our data acquisition can be reliably reproduced using open-source egocentric device platform.

In egocentric vision, neural scene reconstruction has also served as an important building block to enable scene understanding [Gu et al. 2024; Straub et al. 2024; Tschernezki et al. 2023] and contextual AI with human interaction [Lv et al. 2024; Plizzari et al. 2024; Yi et al. 2024]. However, it has been a persistent challenge to acquire high quality reconstructions from egocentric devices in these prior works using existing algorithms. We believe that our work can pave the way for a scalable and accessible high quality neural reconstruction in the rapid growing egocentric research areas.
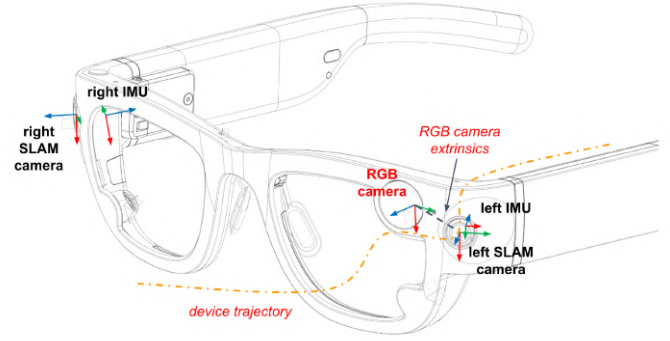

Fig. 2. The layout of sensors for state estimation in a Project Aria device. The device trajectory is represented at high frequency at IMU rate. The red color highlighted the input information used within Gaussian Splatting reconstruction.

3 Background

One of the key distinguishing characteristic we present in this paper is to **focus on reconstructing a video captured by an egocentric form factor device**. This is different from commonly used reconstruction datasets composed of static image snapshots using phones [Barron et al. 2022; Mildenhall et al. 2019], professional cameras [Barron et al. 2023] or high-end 3D scanners [Knapitsch et al. 2017].

We use the open platform Project Aria [Engel et al. 2023], which is a representative egocentric device with form factor and software for future commodity 3D sensors. We consider the following properties as the essential inputs of our study:

(1) With a high-frequency closed-loop device trajectory at the IMU rate (e.g., 1 kHz for Project Aria), we can approximate the 6DoF poses as a piecewise continuous function with respect to timestamps. The raw sensor measurements are timestamped on a common clock at nanosecond resolution. With the state estimation as described in Section 4, we can reliably calculate the asynchronous posed rays using the estimated pixel timestamp derived from first principles.
(2) The device provides a raw sensor output for the RGB camera, including parameters such as gain, exposure, and a calibrated vignette image. This facilitates the modeling the physical image formation process without approximation. Our captured images do not undergo any additional image signal processing (ISP), such as denoising or local tonemapping.

In the following, we will discuss the importance of acquiring high-frequency RGB sensor calibration in Section 4. Then we will introduce a Gaussian Splatting pipeline that leverages the high frequency trajectory and raw sensor models in Section 5. We will further discuss details in capture settings that can improve scene reconstruction in Section 6.

4 Visual Inertial Bundle Adjustment

The state estimation pipeline of an egocentric device (Fig. 3) contains high-frequency (1KHz) device trajectory and online sensor calibration values. In Project Aria device, the sensors (Fig. 2) used for
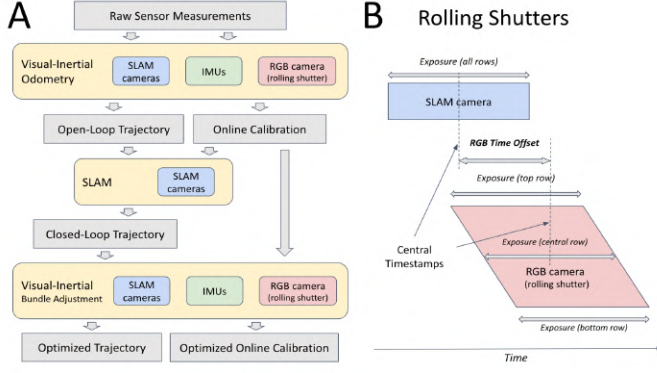
Fig. 3. **A.** An overview of the state estimation pipeline. Among them, the VIBA process handle the rolling-shutter RGB camera in a global bundle adjustment. We provide an exemplification of the rolling shutter properties in **B.**, which are handled in the VIBA step. VIBA models the rolling-shutter RGB camera and outputs accurate timestamps with poses for pixel exposed at different rows during the readout time.

estimation are, additionally to the RGB camera, two monochrome global-shutter SLAM cameras and two inertial motion units (IMU). The full state estimation in our system follows the following steps:

(1) The visual inertial odometry (VIO) system fuses sensor measurements computing an incrementally estimated (open-loop) frame-rate device trajectory and online calibration of all sensors.
(2) The SLAM system uses monochrome global-shutter SLAM cameras to provide loop closures and multi-recording relocalizations, computing a closed-loop trajectory.
(3) Resulting estimate is batch-optimized in a **visual-inertial bundle adjustment (VIBA)** step, which improves the accuracy while maintaining loop-closing constraints. Trajectories formed by all frames, including RGB frames and all sensor calibrations are re-estimated in a joint optimization step.

Step (3) provides the essential calibration for the RGB cameras that our reconstruction system depends on, which we will refer to as *VIBA*. Compared to an alternative off-the-shelf bundle adjustment system such as COLMAP [Schönberger and Frahm 2016], VIBA has a few essential differences:

(1) RGB camera calibration uses a rolling-shutter aware model jointly with global-shutter SLAM cameras.
(2) Camera time offsets are optimized as part of the model when the hardware is unable to provide accurate trigger times.
(3) VIBA re-estimates all calibration parameters at IMU frequency for improved estimate accuracy while simultaneously making reprojection errors sub-pixel on tracked points.
(4) The system scales well to long egocentric videos.

The output of VIBA ensures a high-frequency rolling-shutter aware image formation model, which existing system does not provide. To the best of our knowledge, no existing work investigated the importance of this feature for 3D reconstruction, and the improvement of calibration estimate is especially relevant in the regions that are poor of tracked points, where mismatch might occur if we only tried to improve reprojection errors on the sparse set of

tracked points. The SLAM pipeline, including the step with VIBA, is now accessible through the machine perception service in Project Aria tools. The Project Aria Docs website offers comprehensive instructions on how to access and utilize this tool. We employed the publicly available tool to obtain all input in the paper.

5 Methods

We use Gaussian Splatting (3D-GS) [Kerbl et al. 2023] as the scene representation, a popular framework for efficient photorealistic scene reconstruction. Unlike existing approaches [Moenne-Loccoz et al. 2024; Seiskari et al. 2024], our proposed approach handles camera motions such as rolling shutter, and lens distortion with no change required in the standard Gaussian rasterization and can be applicable to its broad family of advanced variants.

We first discuss a few key notations in 3D-GS and its variants. Then we will illustrate how we update its image rasterization formulation model to handle the common artifacts in egocentric camera sensing. We provide the full implementation details of the preprocessing steps of all input data in the supplementary material.

5.1 Gaussian Splatting

The 3D-GS represents the scene $\mathcal{S}$ as a set of 3D Gaussians $\mathbf{G} = \{\boldsymbol{\mu}, \mathbf{R}, \alpha, \mathbf{c}\}$. Each Gaussian is determined by its 3D mean position $\boldsymbol{\mu} \in \mathbb{R}^3$ and 3D covariance $\Sigma \in \mathbb{R}^{3\times3}$. To approximate a semi-definite 3D covariance, it is parameterized as $\mathbf{RSS}^T\mathbf{R}^T$ using rotation $\mathbf{R} \in SO3$ and scale $\mathbf{S} \in \mathbb{R}^3$. To render an image, all 3D Gaussians are first projected to 2D given the camera pose $\mathbf{T} \in SE3$ as sorted 2D Guassians according to the projected depth value. The pixel color $\mathbf{C}(\mathbf{u})$ is an accumulation of the Gaussian color $\mathbf{c} \in \mathbb{R}^3$ and opacity value $\alpha \in \mathbb{R}^1$ by traversing the list front-to-back. To simplify the notation, we represent the rasterization process to acquire the color of a pixel using the following rendering function:

$$\mathbf{C}(\mathbf{u}) = \pi(\mathbf{u}, \mathcal{S}, \mathbf{T}) \tag{1}$$

The above rasterization function in Eq. 1 can also generalize to other Gaussian alternative, e.g. 2D-GS [Huang et al. 2024], with slight different parameterization of the scene $\mathcal{S}$. We will use this function in Eq. 1 to refer the broad family of Gaussian rasterization approaches in following sections and results.

5.2 Image rasterization model with high frequency poses

We represent the high frequency trajectory as a piecewise continuous function with the 6DoF pose, which support the pose query at any time t as $\mathbf{T}(t) = f_\mathbf{T}(t)$. For a rolling-shutter camera, each row of a pixel has its asynchronous 6DoF pose given the query time at $t(\mathbf{u})$ from from the image capture time $t(\mathbf{0})$ and readout time Δdt_r:

$$t(\mathbf{u}) = t(\mathbf{0}) + \left(\frac{\mathbf{u}_h}{H}\right) \cdot \Delta dt_r \tag{2}$$

where $t(\mathbf{0})$ represent the capture time of the first row pixel for the image with height H and $\mathbf{u}_h$ is its row index.

The physical image formation for each pixel accumulates photons of projected scene irradiance during a fixed exposure time t_e and amplified by an analog or digital gain value g. The image irradiance is further transformed and compressed into an image with certain dynamic range. We can calculate the color of each pixel $\mathbf{C}(\mathbf{u})$ by all
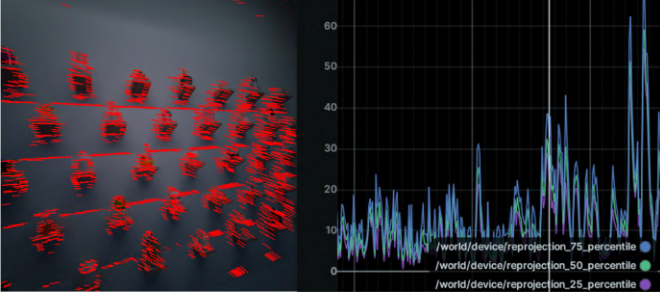
Fig. 4. We visualize the impact of motion during image read-out time. In the left image, we visualize the reprojection vector of sparse scene depth during the image read out time. On the right, we calculated the magnitude of reprojection error for all points per-frame, and plot the 25,50,75 percentile of distribution along this trajectory in time. In this particular frame, 50 percentile of the points have about reprojection error of 30. Such errors will cause misalignment in reconstruction if not handled properly.

the sampled pixels with poses $\mathbf{T}$ starting $t(\mathbf{u})$ within the sampled exposure interval t_e, as the following updated function:

$$\mathbf{C}(\mathbf{u}) = \phi(\omega(\mathbf{u}) \int_0^{t_e} \pi(\mathbf{u}, \mathcal{S}, \mathbf{T}(t(\mathbf{u}) + t))dt) \tag{3}$$

where $\omega(\mathbf{u})$ is a per-pixel linear weight that combines the effect of analog gain g, lens shading $V(\mathbf{u})$ and normalization factor. $\phi(\cdot)$ represents the camera response function. We will discuss a few importance factors in practice as following.

Rolling-shutter with lens distortion: The query time in Eq. 2 is calculated based on the raw undistorted image. After image rectification, the linear relationship for each pixel respect to their row number will not hold which prohibits existing solution [Seiskari et al. 2024] to be applicable. We propose to generate a index ratio as a look-up table for the rectified image $\mathcal{R}(\mathbf{u})$, where each pixel value is a ratio representing its row number relative to the image height. We adjust Eq. 2 to the following time query:

$$t(\mathbf{u}) = t(0) + \mathcal{R}(\mathbf{u})\Delta dt_r \tag{4}$$

Camera motion sampling: A discrete form of Eq. 3 requires sufficiently sampling the possible motions that could result in substantial pixel offsets during the exposure or readout time. We use the scene depth to estimate the reprojection error during a temporal bracket. Specifically, the sparse depth value is triangulated from the tracked points in global shutter SLAM cameras and serve as the 3D anchors for camera motion estimation. We estimate the length of temporal bracket that ensures half of the reprojected pixels have fewer than 1 pixel reprojection error. As a result, the full-resolution RGB camera with a readout time of approximately 16 ms has an average of eight motion samples within the readout time.

The importance of motion sampling: The human head is in constant motion and can achieve rotational velocities of several hundred degrees per second. Fig. 4 shows the artifact of pixel reprojection errors during the image readout time. In our ablation study, we will show correcting modeling the pixel motions can drastically improve the reconstruction quality.

Rasterization: We batch rasterize the image based on the number of samples in $\mathcal{R}(\mathbf{u})$ during the forward process and then use a gather operation to synthesize a final image. The number of pixels contribute to the backward gradients are same as the single image. We employ the aforementioned camera motion sampling strategy to determine the sample bracket within $\mathcal{R}(\mathbf{u})$. For a quasi-static viewpoint, only one pose sample is needed, whereas 8-16 samples may be required for a fast-moving view.

This approach is general to camera models and can be particular helpful for those containing high-order distortions. In our example, the RGB camera is a fisheye model with high order of coefficients which no existing rasterization implementation supports. Unlike existing approach that require customized rasterization kernel [Seiskari et al. 2024] to camera with particular calibration or using raytracing [Condor et al. 2024], our proposed rasterization in Eq. 3 with rectified index look up table require no change to general 3D-GS or its variants such as 2D-GS [Huang et al. 2024] using common pinhole or fisheye rectified images [Liao et al. 2024].

5.3 Additional factors

Preserving dynamic range of the scene: Different from datasets that often estimate the scene radiance directly using tonemapped images, the Gaussian color $\mathbf{c}$ of the scene $\mathcal{S}$ in Eq. 3 is in linear space by default, which can hurt the optimization particular in high dynamic range scenarios. Thus, we explicitly encode scene color as a gamma compressed scene irradiance $\mathbf{c} = \phi^{-1}(\mathbf{r})$. We use the same notation $\phi(\cdot)$ to represent the inverse of camera response for simplicity. We use gamma value 2.2 as our default setting.

Image blur: For an auto-exposed camera, the exposure time is related to the auto-exposure target and the scene luminance. In bright outdoor scenarios, where the exposure of 0.5 ms is sufficient, the amount of motion during the exposure time is limited. However, in indoor scenes, the exposure can be ten times longer, leading to significant motion blur. Although Eq. 3 models the blur formulation and can be used to optimize deblurred pixels as existing work [Seiskari et al. 2024], we find that it is challenging to reliably reconstruct sharp details from the blurry image without a very dense capture, which is hard to be fulfilled by human movement in practice. Instead, we can record less blurry images with shorter exposure time, which motivated our capture process to be discussed in Section 6.

Handling noise: The Project Aria device uses analog gain to compensate image brightness in low light scenario resulting in strong photon shot noise that affects the image quality. For a gamma compressed image, an approximated square-root gamma value can whiten this photon shot noise and can be handled by the L1/L2 reconstruction loss without additional efforts [Lehtinen et al. 2018]. Different from existing HDR reconstruction [Mildenhall et al. 2022; Singh et al. 2024] that target a HDR recovery using special losses for extreme dark scenario, we found the standard 3D-GS training objective is effective handle such indoor scenarios.

6 Capturing Egocentric Video

We record egocentric videos as 10 FPS JPEG-compressed 8MP images using Project Aria [Engel et al. 2023]. No denoising or deblurring

is applied to the input video. We rectified the images to 2400x2400 resolution and use them for training.

In indoor scenario with illumination under 300 lux, standard auto-exposure system will increasing either the exposure time or the analog gain. Increasing the exposure time can result in motion-blurred captures, while increasing the gain can introduce more noises. In Section 5.3, we discuss certain level of photon shot noise can be handled via a noise-to-noise reconstruction process while it remains challenging to handle blur. This inspired us to use a special capture treatment to handle low light scenarios.

Capturing videos in indoor environment. Inspired by high dynamic range (HDR) image processing algorithms that recover HDR images from multiple fast exposure captures [Hasinoff et al. 2016], we propose capturing videos using fast exposures to minimize the impact of motion blur and recover high dynamic range, noise-free 3D scenes from the 3D reconstruction. We limit the RGB camera exposure time to a maximum of 2 ms. This works for general indoor environments measured with illuminations between 150-300 lux. Low light indoor scenarios is extremely challenging for egocentric video which is a limitation we leave for future work.

Aria scene evaluation dataset. We collected two categories of egocentric recordings to evaluate the algorithm in diverse conditions. Each has 6 recordings in scenes with varying complexity. The *outdoor recordings* have high dynamic range illuminated with a minimum of 3K lux. We used auto-exposure with varying exposures and minimum analog gain. Little motion blur or noise is present in these recordings. We use them to evaluate free viewpoint reconstruction in unconstrained large-scale environments. For *indoors recordings*, we collect them with the proposed indoor capture protocol. Fig. 6 and Fig. 7 contain some visual examples and evaluation. We include more details of the dataset in supplementary materials.

The full processing of the dataset is based on tools from the open-sourced Project Aria platform. We release the collected datasets as examples and we hope it can help the community can bring the learnings to various applications at a bigger scale.

7 Experiments

Baselines and ablations. We use 3D-GS in GSplats [Ye et al. 2024] as our main rasterization method to represent the family of Gaussian Splatting algorithms. For all comparisons, we use the same hyper-parameters following the baseline. We also include results that use 2D-GS as the rasterization method, which demonstrates our method is applicable to other Gaussian variants. We provide the following evaluation.

(1) **Splatfacto.** The Splatfacto in Nerfstudio integrates a few advanced features of Gaussian Splats. We use the same GSplats version (1.4) in our training and this baseline for a fair comparison. We provide the same calibration for RGB camera that for our ablations that does not use VIBA, which represents the most common reconstruction baseline in existing work [Gu et al. 2024; Yi et al. 2024] using egocentric devices.

(2) **3DGS-on-move [Seiskari et al. 2024].** This represents the state-of-the-art work that similarly rasterizes the Gaussians with an explicit image formation model. It rasterizes the Gaussians

from a moving camera using the camera velocity information. In addition, it jointly optimizes the camera parameters during training to correct potential pose errors. We calculated the RGB camera velocity and provide them to this algorithm as initialization. Other than this, it use the same input as Splatfacto. We use the default parameters setting and the same training iterations as all other baselines.

(3) **Ours Egocentric GS (3D-GS/2D-GS).** This is our implementation using the extensions we discussed in Section 5.2, based on closed-loop device trajectories and online calibrations. We use 3D-GS as the default choice of Gaussian rasterization pipeline if not clarified.

(4) **Ablations: without VIBA.** We use the high frequency closed-loop trajectory calculated using the SLAM cameras and the factory calibration for RGB camera before VIBA. As splatfacto, it represents the commonly used setting in previous work.

(5) **Ablations: without motion sampling.** We disable motion aware pose sampling technique in Section 5 to compensate the rolling shutter effect. As all other baselines, we only use the center row pose represent the image pose, calculated from the VIBA calibrated trajectory.

(6) **Ablations: without scene gamma.** In Eq. 3, we represent scene radiance in linear space with no gamma conversion.

Evaluation datasets. We evaluate our algorithm using the following data:

(1) **Aria scene dataset.** To evaluate the algorithm performing at scale, we collected a set of indoor and outdoor egocentric videos using the protocols discussed in Section 6.

(2) **Digital Twin Catelog (DTC) dataset.** We use the egocentric recordings within the Digital Twin Catalog dataset [Dong et al. 2025], which has precisely 3D aligned ground truth. These videos are recorded with the same lab lighting condition and fixed exposure gain inputs using Project Aria device. We evaluate the predicted geometry quality using the depth and normal rendered from the 3D ground truth.

(3) **Quest scene dataset.** We further collect one sequence using the Meta Quest 3 device. We process the dataset following the same proposed process. This evaluation demonstrate the generalization of the proposed method to cope with other data recorded from other commodity egocentric headsets.

For all of the evaluations, we create the validation set following the common practices as [Mildenhall et al. 2019] that held out every 8th image as the validation images and use the rest for training. We perform all evaluations using PSNR and SSIM as the main image quality metrics.

Results. Table 1 demonstrates the quantitative evaluations on Aria scene dataset, in both outdoor and indoor recordings, which includes the comparison to the baselines and ablations. Table 2 provides the comparison on the DTC dataset. Fig. 6 show a qualitative comparisons of our method to baselines. Fig. 7 compares our ablation settings. Fig. 1 highlights the high dynamic range noise-free reconstruction in indoor environment using fast exposure captures. We summarize the findings in the following.

Table 1. Quantitative evaluations for the Aria scene dataset. We separate the scenarios for outdoor and indoor scenarios. Splatfacto refers to [Ye et al. 2024] and 3DGS-on-move refers to [Seiskari et al. 2024].

Ourdoor scenes	bike shop		steakhouse patio		pop-up shop		sunroom		garden		restaurant patio	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
splatfacto	26.98	0.803	26.30	0.777	26.48	0.766	22.82	0.735	20.32	0.647	20.34	0.701
3DGS-on-move	27.07	0.806	25.80	0.774	26.91	0.773	23.33	0.752	19.68	0.644	20.33	0.706
Our Egocentric-GS	**29.98**	**0.838**	**28.38**	**0.805**	**29.03**	**0.797**	**27.03**	**0.788**	**24.00**	**0.704**	**25.30**	**0.787**
w/o VIBA	27.68	0.800	27.03	0.782	27.64	0.771	25.22	0.751	22.58	0.670	22.32	0.725
w/o motion sampling	28.83	0.819	27.70	0.792	28.29	0.783	26.15	0.766	22.90	0.678	23.31	0.746
w/o scene gamma	29.04	0.836	27.21	0.790	28.16	0.795	21.76	0.746	21.74	0.685	22.52	0.765
Indoor scenes	Library		Plant hallway		Open hallway		Micro Kitchen		Multi-Floor		Livingroom	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
splatfacto	23.62	0.522	21.65	0.723	23.85	0.530	23.05	0.547	24.92	0.520	27.41	**0.546**
3DGS-on-move	22.03	0.501	20.449	0.712	21.85	0.511	23.35	0.544	23.99	0.511	27.62	**0.546**
Our Egocentric-GS	**24.59**	**0.544**	**23.97**	**0.765**	**26.61**	**0.554**	**27.11**	**0.579**	**25.96**	**0.535**	**27.73**	0.546
w/o VIBA	23.91	0.525	22.64	0.731	25.63	0.539	25.54	0.559	25.26	0.522	26.35	0.527
w/o motion sampling	23.98	0.525	22.82	0.740	25.66	0.540	25.94	0.565	25.57	0.527	27.28	0.535
w/o scene gamma	24.22	0.542	21.02	0.744	25.01	0.547	24.48	0.561	25.66	**0.535**	27.33	0.545

Table 2. Quantitative evaluation of appearance and geometry reconstruction on egocentric DTC dataset. [Dong et al. 2025]. We evaluate depth using scale-invariant L1 loss and evaluate normal using L1 loss.

	PSNR ↑	Depth ↓	Normal ↓
Ours (3D-GS)	**29.83**	**0.1505**	**0.3078**
w/o motion sampling	28.93	0.1769	0.3171
w/o VIBA	28.52	0.1730	0.3274
w/o scene gamma	28.76	0.1768	0.3236
splatfacto [Ye et al. 2024]	24.81	0.1749	0.6627
Ours 2D-GS	**29.54**	**0.1474**	**0.1509**
w/o motion sampling	28.75	0.1755	0.2112

Table 3. Quantative evaluations on Meta Quest 3 recording dataset.

	PSNR ↑	SSIM ↑	LPIPS ↓
Ours (3D-GS)	**29.54**	**0.9147**	**0.2271**
w/o VIBA	27.27	0.8737	0.2731
w/o motion sampling	28.85	0.9012	0.2344

Comparison to existing work. Compared to both splatfacto and 3DGS-on-move, our method significantly outperform them in both quantitative and qualitative comparisons. Both baselines fail to recover scene details and create significant floaters in the scene. Our solution without using VIBA or camera motion sampling also outperforms them in most scenes. Although 3DGS-on-move rasterizes a physical image model considering both rolling-shutter and blur, its joint optimization with Gaussians splatting does not necessarily provide better visual quality compared to the splatfacto. Both methods come close in a small dense captured scene (Livingroom), but we observe larger performance gap for scenes with increasing scale and complexity.

The effect of VIBA and motion sampling. We observe a big improvement in visual quality (2-3db in PSNR) when using optimized calibration from VIBA, and we can also observe consistent improvement (1db in PSNR) when using the motion sampling to compensate the camera rolling-shutter effect. In DTC evaluation, we can observe

using VIBA and motion sampling in color will also contribute to better geometry modeling.

The effect of modeling gamma compressed scene radiance. Explicitly modeling the radiance in gamma space can boost visual quality in all scenes and have more significantly impact in outdoor scenes. As mentioned in Section 5, it helps to model scenes with higher dynamic range.

Generalization to other headset. We apply the same process to one recording using Meta Quest 3 dataset and report the quantitative comparison in Table 3 and perform the ablation study on 3D-GS baselines. Similar as the observation in Aria datasets, we can consistently observe the performance improvement measured in all metrics when using VIBA and motion sampling.

High dynamic range rendering. Our reconstruction in low-light indoor scenario preserve the dynamic range of the scene, which can produce enhanced rendering after reconstruction. In Fig. 1, we show a visual example to simulate a rendering camera with 3x gain. We can see the improved clarity of details and text despite the input video is noisy and dark.

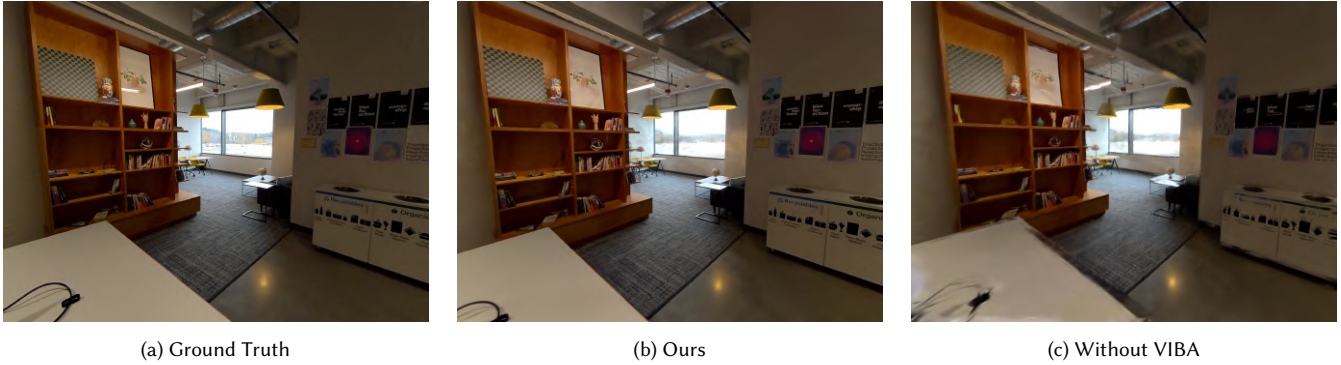| (a) Ground Truth | (b) Ours | (c) Without VIBA |

Fig. 5. Qualitative comparison on a Quest 3 device. We can reconstruct scenes with sharper details using VIBA.

8 Conclusion

In this paper, we describe a system for capturing videos using an egocentric device and a method for reconstructing photorealistic scenes. We argue the importance of correctly calibrating and modeling the physical image formation model from first principles. In our evaluation, our results produce high quality rendering with sharp details while existing methods fail to do so. The proposed solution is built on an open-source egocentric glasses platform, and we tested it across scenes with varying complexity and lighting conditions. We further validated our approach on a commercial headset, leading to consistent conclusions across different platforms. With the rising demand for lightweight egocentric glasses, we believe our method can benefit various applications and inspire future methods to design better scene reconstruction algorithm and hardware end-to-end.

Limitations and future work. Firstly, while our method surpasses current state-of-the-art techniques in scene reconstruction, reconstructing any scene from any human trajectory remains an unsolved challenge. Unlike static captures, egocentric video may lack sufficient view coverage, leading to reconstruction artifacts. Recent methods using sparse views show promise and could inform future improvements. Secondly, like most reconstruction algorithms, we assume a static scene, which is a significant limitation. Human body parts, shadows, illumination changes, and scene motions are often present and should be addressed in future work to enhance scalability. Lastly, our system struggles in extremely low light conditions (<50 lux), resulting in insufficient signal capture and failure. Future advancements in image sensing and reconstruction algorithms may better address this issue.

Acknowledgments

References

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*. 5470–5479.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2023. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. *ICCV* (2023).

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to See in the Dark. In *CVPR*.

Jorge Condor, Sebastien Speierer, Lukas Bode, Aljaz Bozic, Simon Green, Piotr Didyk, and Adrian Jarabo. 2024. Don't Splat your Gaussians: Volumetric Ray-Traced Primitives for Modeling and Rendering Scattering and Emissive Media. arXiv:2405.15425 [cs.GR] https://arxiv.org/abs/2405.15425

Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, Sean Christofferson, James Fort, Xiaqing Pan, Mingfei Yan, Jiajun Wu, Carl Yuheng Ren, and Richard Newcombe. 2025. Digital Twin Catalog: A Large-Scale Photorealistic 3D Object Digital Twin Dataset. In *CVPR*.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. *arXiv preprint arXiv:2308.13561* (2023).

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields Without Neural Networks. In *CVPR*. 5501–5510.

Michael Goesele, Daniel Andersen, Yujia Chen, Simon Green, Eddy Ilg, Chao Li, Johnson Liu, Grace Kuo, Logan Wan, and Richard Newcombe. 2025. Imaging for All-Day Wearable Smart Glasses. arXiv:2504.13060 [cs.CV] https://arxiv.org/abs/2504.13060

Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. 2024. EgoLifter: Open-world 3D Segmentation for Egocentric Perception. *ECCV* (2024).

Sam Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *SIGGRAPH Asia* (2016). http://www.hdrplusdata.org/hdrplus.pdf

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery. doi:10.1145/3641519.3657428

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG* 42, 4 (2023), 1–14.

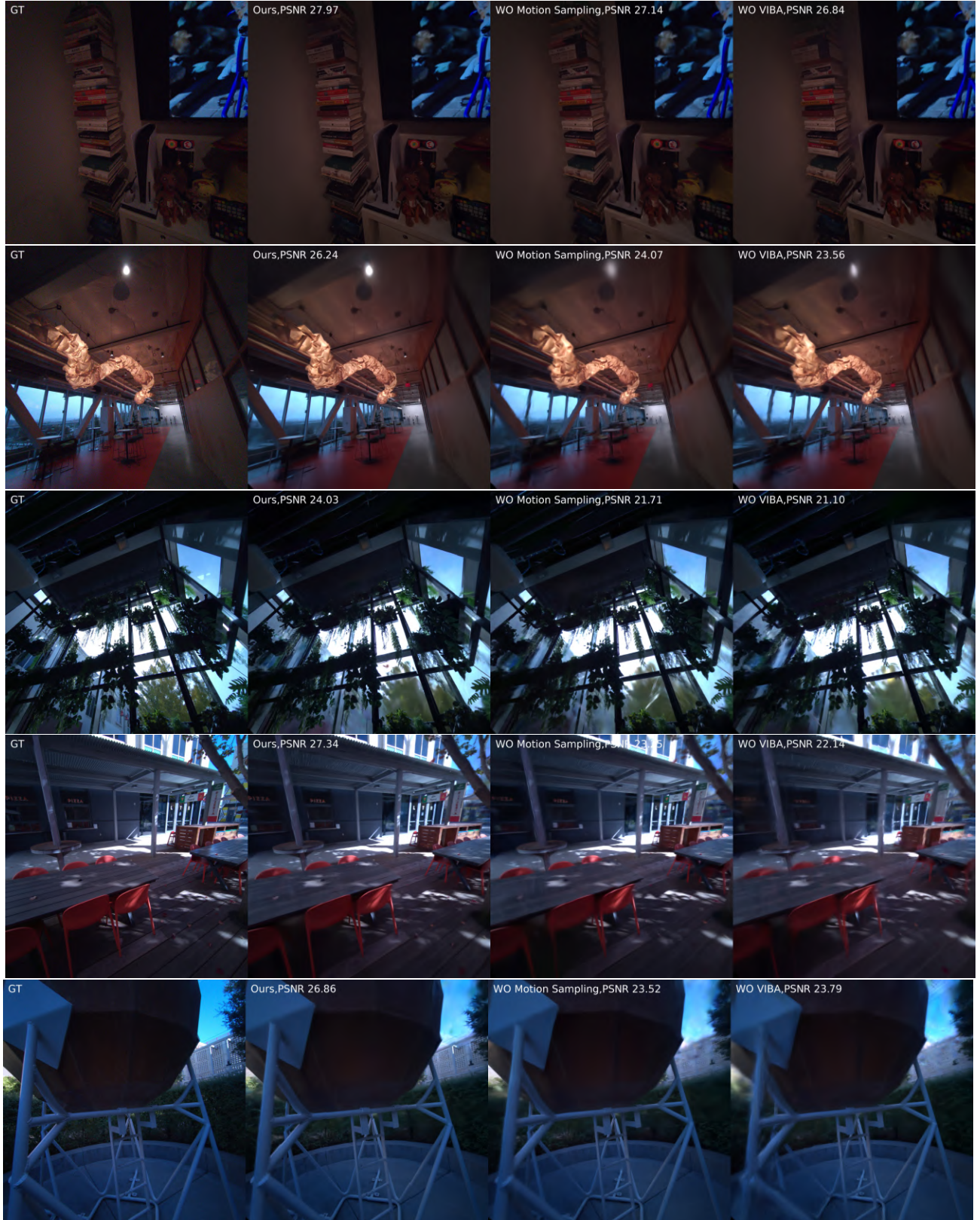Fig. 6. Qualitative comparisons to baseline approaches Splatfacto and 3DGS-on-move.

Fig. 7. Qualitative comparisons of ablations. Better visualized in full resolution.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM TOG* 36, 4 (2017).

Byeonghyeon Lee, Howoong Lee, Xiangyu Sun, Usman Ali, and Eunbyung Park. 2024. Deblurring 3D Gaussian Splatting. arXiv:2401.00834 [cs.CV]

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. 2018. Noise2Noise: Learning Image Restoration without Clean Data. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2965–2974. https://proceedings.mlr.press/v80/lehtinen18a.html

Tianye Li, Mira Slavcheva, Michael Zollhoefer, 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 *CVPR*. 5521–5531.

Zimu Liao, Siyan Chen, Rong Fu, Yi Wang, Zhongling Su, Hao Luo, Li Ma, Linning Xu, Bo Dai, Hengjie Li, Zhilin Pei, and Xingcheng Zhang. 2024. Fisheye-GS: Lightweight and Extensible Gaussian Splatting Module for Fisheye Cameras. arXiv:2409.04751 [cs.CV] https://arxiv.org/abs/2409.04751

Jinsong Liu, Michael S. Brown, and David Suter. 2010. Richardson-Lucy deblurring for scenes under a projective motion path. *IEEE TPAMI* 32, 12 (2010), 2228–2239. doi:10.1109/TPAMI.2010.222

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. 2024. Aria Everyday Activities Dataset. arXiv:2402.13349 [cs.CV] https://arxiv.org/abs/2402.13349

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. 2021. Deblur-NeRF: Neural Radiance Fields from Blurry Images. *arXiv preprint arXiv:2111.14292* (2021).

Alexander Mai, Peter Hedman, George Kopanas, Dor Verbin, David Futschik, Qiangeng Xu, Falko Kuester, Jon Barron, and Yinda Zhang. 2024. EVER: Exact Volumetric Ellipsoid Rendering for Real-time View Synthesis. arXiv:2410.01804 [cs.CV] https://arxiv.org/abs/2410.01804

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. 2022. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. *CVPR* (2022).

Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM TOG* 38, 4 (2019), 1–14.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Nicolas Moenne-Loccoz, Ashkan Mirzaei, Or Perel, Riccardo de Lutio, Janick Martinez Esturo, Gavriel State, Sanja Fidler, Nicholas Sharp, and Zan Gojcic. 2024. 3D Gaussian Ray Tracing: Fast Tracing of Particle Scenes. *ACM Transactions on Graphics and SIGGRAPH Asia* (2024).

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG* 41, 4 (2022), 1–15.

Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE international symposium on mixed and augmented reality*. IEEE, 127–136.

Muyao Niu, Tong Chen, Yifan Zhan, Zhuoxiao Li, Xiang Ji, and Yinqiang Zheng. 2024. RS-NeRF: Neural Radiance Fields from Rolling Shutter Images. In *ECCV*.

Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. 2024. Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. In *ArXiv*.

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *CVPR*.

Otto Seiskari, Jerry Ylilammi, Valtteri Kaatrasalo, Pekka Rantalankila, Matias Turkulainen, Juho Kannala, and Arno Solin. 2024. Gaussian Splatting on the Move: Blur and Rolling Shutter Compensation for Natural Camera Motion. In *ECCV*.

Shreyas Singh, Aryan Garg, and Kaushik Mitra. 2024. HDRSplat: Gaussian Splatting for High Dynamic Range 3D Scene Reconstruction from Raw Images. In *BMVC*.

Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. 2024. EFM3D: A Benchmark for Measuring Progress Towards 3D Egocentric Foundation Models. arXiv:2406.10224 [cs.CV] https://arxiv.org/abs/2406.10224

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797* (2019).

Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. 2023. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. 2023. BAD-NeRF: Bundle Adjusted Deblur Neural Radiance Fields. In *CVPR*. 4170–4179.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. 2024. gsplat: An Open-Source Library for Gaussian Splatting. *arXiv preprint arXiv:2409.06765* (2024). arXiv:2409.06765 [cs.CV] https://arxiv.org/abs/2409.06765

Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. 2024. Estimating Body and Hand Motion in an Ego-sensed World. *arXiv preprint arXiv:2410.03665* (2024).

Zehao Yu, Torsten Sattler, and Andreas Geiger. 2024. Gaussian Opacity Fields: Efficient Adaptive Surface Reconstruction in Unbounded Scenes. *ACM Transactions on Graphics* (2024).

Lingzhe Zhao, Peng Wang, and Peidong Liu. 2024. BAD-Gaussians: Bundle Adjusted Deblur Gaussian Splatting. In *ECCV*.