OmniLife360: A Benchmark for 3D Reconstruction from In-the-Wild 360° Captures

Zonglin Zhao*, Bowen Zhang*, Yatai Li*, Chao Liang, and Zhipeng Zhang†

AutoLab, School of Artificial Intelligence, Shanghai Jiao Tong University
`https://github.com/AutoLab-SAI-SJTU/Omni4DGS`
* Equal contribution. † Corresponding author

Abstract. The widespread adoption of consumer 360° cameras has created a growing demand for robust in-the-wild panoramic 3D reconstruction. Existing methods remain fragile on long-horizon trajectories with strong ego-motion and dynamic interference. Quantifying this gap at scale is challenging because available benchmarks are often short, constrained, or captured in controlled settings, limiting evaluation and progress. To address this, we introduce **OmniLife360**, a benchmark for 3D reconstruction from in-the-wild 360° captures. It spans diverse scenes, environments, and motion dynamics captured by individuals across real-world activities, enabling rigorous evaluation of reconstruction robustness in complex natural conditions. Extensive evaluations reveal that state-of-the-art methods excel in controlled settings but suffer significant degradation in dynamic scenarios. Motivated by these findings, we propose **Omni4DGS**, a motion-aware Gaussian Splatting framework that combines motion decomposition with lightweight supervision to better handle motion variations, dynamic distractors, and long-horizon pose drift. Experiments show that Omni4DGS consistently improves reconstruction quality and stability in dynamic 360° scenarios, reducing motion-induced artifacts and producing cleaner renderings under strong scene motion. The benchmark is available at `https://huggingface.co/datasets/AutoLab-SJTU/OmniLife360`.

Keywords: Panoramic 3D Reconstruction · In-the-Wild 360° Benchmark · 4D Gaussian Splatting

1 Introduction

The widespread adoption of commercial 360° cameras is progressively reshaping how people capture and share visual experiences. Omnidirectional videos have become increasingly prevalent in user-generated content across daily life logging, travel recording, and outdoor sports. Beyond passive viewing, there is a growing demand to transform these raw panoramic recordings into persistent and navigable 3D experiences through immersive scene reconstruction and novel view synthesis. Compared with traditional narrow field-of-view imagery, 360° cameras capture continuous and globally consistent spatial context within a single exposure. This advantage alleviates viewpoint discontinuities and avoids complex stitching pipelines. However, reconstructing accurate 3D geometry from unconstrained panoramic videos remains highly challenging due to spherical projection distortions, unpredictable camera movements, cumulative trajectory drift, and pervasive dynamic interference.

Current panoramic reconstruction methods generally fall into two distinct paradigms, yet both exhibit notable limitations. Scene-specific optimization approaches [1,16,23] achieve high fidelity on training views but struggle with novel perspective extrapolation. Their lengthy optimization times, typically requiring tens of minutes per scene, also hinder practical deployment. Alternatively, generalizable feed-forward methods [5, 17, 21, 28] offer rapid inference from sparse inputs yet suffer severe degradation under untrained scenes and distinct perspective changes. More crucially, both paradigms struggle to accurately model moving elements such as pedestrians, vehicles, or the camera operator. This difficulty arises from extreme spherical projection distortions, unpredictable camera motions, and pervasive dynamic interferences, ultimately resulting in obvious rendering artifacts and blur in dynamic regions.

The evaluation and advancement of robust solutions are fundamentally limited by the shortcomings of existing panoramic benchmarks. Most datasets are confined to restricted settings. Synthetic or scanned indoor environments like OmniBlender [6], HM3D [20], and Replica [25] lack natural camera motion, making it difficult to replicate authentic motion blur and photometric inconsistencies. Meanwhile, real-world benchmarks [2, 9, 11, 14] are typically limited to extremely short clips lasting only a few seconds, or rely on heavy and highly specialized vehicular rigs that yield smooth and constrained trajectories. Consequently, the community lacks a comprehensive testbed to evaluate the erratic camera movements, long-term cumulative drift, and unconstrained dynamics inherent to spontaneous and daily life recordings.

To fill this critical gap, we introduce **OmniLife360** as a novel large-scale benchmark for evaluating extended panoramic reconstruction in real-life and behavior-driven capture scenarios. Constructed through a systematic human and AI collaborative pipeline, OmniLife360 contains 2407 high-quality and minute-level panoramic video sequences, totaling roughly 66 hours of footage at 4K resolution. Unlike previous datasets, OmniLife360 strictly ensures geometric validity through geometry-aware temporal slicing, while preserving high semantic diversity and dynamic complexity. It encompasses 64 distinct scene categories and 31 action classes. This authentically introduces the complex illumination, rapid camera rotations, and drastic movements that define the ultimate challenges of user-generated content.

Through extensive evaluation of leading methods on OmniLife360, we uncover two fundamental weaknesses including significant cumulative drift along extended trajectories induced by strong camera motion, and pronounced confusion between static and dynamic elements caused by pervasive scene dynamics. To address these limitations, we propose **Omni4DGS** as a novel 4D Gaussian Splatting framework specifically designed for unconstrained 360° panoramic videos. Omni4DGS introduces a temporally and spherically aware motion formulation to stabilize long-sequence camera trajectories. It further introduces a panorama-specific velocity decomposition into radial and tangential components with latitude-aware adaptive weighting, improving stability under unit-sphere projection. Together with a dynamic and static decoupling mechanism supervised by 2D masks to explicitly model moving objects, this design substantially mitigates rendering artifacts in unconstrained scenes, enables more accurate dynamic object reconstruction, and establishes a robust baseline for dynamic panoramic reconstruction. Experiments on OmniLife360 show that Omni4DGS achieves clear PSNR gains and markedly lower LPIPS than strong baselines, while suppressing artifacts in motion-affected regions and producing cleaner, more stable reconstructions.

Our main contributions are summarized as follows.

- We introduce **OmniLife360** as the first large-scale benchmark for extended 360° panoramic reconstruction in real-life capture scenarios, featuring strong camera motion and scene dynamics.
- We conduct a comprehensive evaluation of existing advanced methods on OmniLife360, revealing fundamental limitations in handling long-term drift and dynamic scene interference.
- We propose **Omni4DGS** as the first 4D Gaussian Splatting framework explicitly tailored for spherical panoramic videos with dynamic and static decoupling, enabling robust reconstruction in complex unconstrained scenarios.

2 Related Work

2.1 Panoramic Datasets and Benchmarks for 3D Reconstruction

Compared with narrow FOV representations, panoramic data provides substantially richer spatial context for continuous scene understanding. However, existing datasets fall into three dominant paradigms, leaving the evaluation of dynamic and long-horizon 3D reconstruction insufficiently explored. Foundational scanned datasets such as HM3D [20] and Replica [25] offer high-resolution digital twins, but they are predominantly limited to static indoor environments and therefore lack pervasive dynamic interference. Synthetic benchmarks including Structured3D [30] and OmniBlender [6] simulate large-scale trajectories, yet they suffer from a pronounced domain gap and cannot faithfully reproduce real-world photometric inconsistencies or natural motion blur. Recent real-world panoramic collections attempt to incorporate dynamic content, but they exhibit notable limitations. Casual captures such as OmniPhotos [2] consist only of very short clips lasting a few seconds. Larger-scale efforts either provide limited environmental diversity, as in the case of 360Loc [10] with only four scene types, or rely on specialized vehicular platforms. For example, KITTI-360 [18] uses a vehicle-mounted multi-sensor rig, producing smooth and constrained motion that misses irregular ego-motion and in-the-wild complexity. As summarized in Tab. 1, existing panoramic datasets are either static, short, synthetic, or captured under

Table 1: Comparison of commonly used datasets and benchmarks for panoramic 3D reconstruction. Dyn. denotes scene dynamic complexity, qualitatively categorized as Static, Low, Med, or High by dynamic objects prevalence and motion intensity.

Dataset	Source	Env.	Res.	#Scenes	Dyn.
HM3D [20]	Scanned	Indoor	≥2K	1,000	Static
Replica [25]	Scanned	Indoor	4K	18	Static
OmniBlender [6]	Synthetic	Both	2K	11	Static
OmniPhotos [2]	Real	Both	5.7K	31	Low
360Roam [9]	Real	Indoor	6K	10	Low
OmniScenes [14]	Real	Indoor	2K	7	Med
360Loc [10]	Real	Both	6K	4	Med
OmniLife360 (Ours)	**Real**	**Both**	**≥4K**	**2,407**	**High**

constrained settings, leaving dynamic and long-horizon reconstruction insufficiently explored. To address these shortcomings, we introduce **OmniLife360**. Built exclusively from spontaneous, user-generated in-the-wild videos, it integrates minute-level temporal continuity, extensive behavioral diversity, and strict geometric validity within a unified benchmark.

2.2 Panoramic 3D Reconstruction Methods

Panoramic 3D reconstruction is challenging due to spherical projection distortions and camera-induced geometric inconsistencies. Although NeRF-based approaches [4,6,7,15] can produce high-quality results, their per-scene optimization cost has motivated a shift toward 3D Gaussian Splatting (3DGS) [13]. Recent panoramic 3DGS methods mainly enhance spherical adaptation through optimization and are typically developed under a static-scene assumption. ODGS [16] reduces distortion with tangent-plane rasterization, and 360-GS [1] further incorporates layout guidance. SC-OmniGS [8] and Seam360GS [23] improve camera modeling via differentiable self-calibration and dual-fisheye calibration, alleviating stitching artifacts. To improve efficiency, feed-forward models such as Splatter-360 [5], OmniSplat [17], and PanSplat [28] employ spherical cost volumes, Yin–Yang grids [12], and multi-scale Gaussian representations for fast synthesis. PanoSplatt3R [21] leverages perspective pretraining to better handle unposed wide-baseline panoramas. However, everyday in-the-wild captures often involve erratic ego-motion, long trajectories, and pervasive dynamics as in OmniLife360, where the static-scene assumption breaks and causes static–dynamic confusion and motion-induced artifacts. To address these limitations, we propose **Omni4DGS**, a 4DGS framework tailored for panoramic 4D reconstruction. Instead of directly extending existing Cartesian-space 4DGS methods, Omni4DGS adopts a spherical-aware formulation with dedicated motion parameterization and latitude-aware adaptation to explicitly handle ERP distortions, resulting in more robust dynamic reconstruction with fewer rendering artifacts under challenging real-world captures.

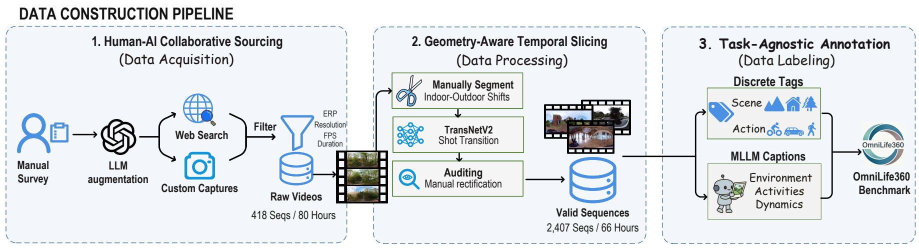

Fig. 1: Overview of the OmniLife360 data construction pipeline. Raw panoramic videos are collected via human-AI collaborative sourcing, refined through geometry-aware temporal slicing to ensure multi-view consistency, and enriched with task-agnostic semantic annotations to form the final benchmark.

3 The OmniLife360 Benchmark

The widespread use of consumer 360° cameras has created strong demand for reconstructing navigable 3D experiences from unconstrained panoramic videos, yet existing panoramic benchmarks remain misaligned with this deployment setting and do not adequately stress long-horizon stability or robustness under real-world dynamics. To address this gap, we introduce **OmniLife360**, a benchmark designed for 3D reconstruction from unconstrained 360° captures. Constructed through a three-stage pipeline (Fig. 1), OmniLife360 combines panoramic modality, extended temporal horizons, and real-world dynamic complexity under a unified evaluation framework. By preserving authentic in-the-wild characteristics while enforcing strict geometric consistency, our benchmark enables rigorous assessment of long-horizon stability and dynamic modeling robustness in realistic scenarios. Representative qualitative examples of the scene diversity, ego-motion regimes, and dynamic interference in OmniLife360 are shown in Fig. 2.

3.1 Human-AI Collaborative Data Collection

To capture authentic "OmniLife" dynamics, we eschewed laboratory setups in favor of a Human-AI collaborative collection protocol. We first conducted a manual survey of user-uploaded panoramic videos across major public platforms to compile a seed set of representative capturing behaviors and intents. These seeds were then used to prompt a Large Language Model (LLM) agent to expand the taxonomy of environments and daily activities, mitigating human search bias and improving scenario coverage.

Guided by this expanded landscape, we comprehensively queried public internet repositories. Candidate videos were required to satisfy several criteria. We retained only native ERP videos with a minimum resolution of 4K to maintain sufficient angular resolution under spherical projection, a frame rate of at least 30 fps to ensure adequate temporal sampling under rapid ego-motion, and a raw

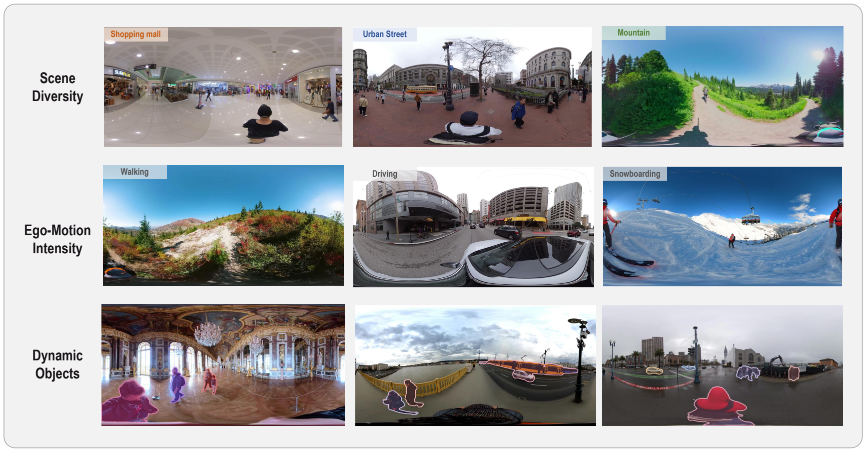

Fig. 2: Qualitative examples from OmniLife360. Row 1 shows diverse real-world environments. Row 2 covers ego-motion regimes from walking to high-speed and extreme activities. Row 3 highlights pervasive dynamic interference, where dynamic object masks are overlaid to visualize moving distractors and the regions most prone to motion-induced artifacts.

duration exceeding 30 seconds to enable meaningful evaluation of long-horizon stability and drift behavior.

We avoided aesthetic-based filtering to retain the natural photometric degradations, motion blur, and illumination variations of spontaneous captures. However, videos containing prominent artificial overlays such as watermarks or heavy mosaic masking were excluded, as such artifacts introduce spurious occlusions and high-contrast patterns that can compromise geometric consistency. To complement scenarios that are underrepresented in publicly available web videos, such as campus traversals and night markets, we additionally captured raw panoramic sequences using a consistent acquisition setup. This collection stage yielded 418 raw panoramic sequences totaling approximately 80 hours.

3.2 Geometry-Aware Temporal Slicing

Raw real-world videos often contain edits, abrupt context changes such as indoor-outdoor transitions, and shot transitions that violate multi-view geometric consistency. To extract geometrically valid, temporally continuous segments, we design a multi-stage temporal slicing pipeline.

First, human annotators manually segmented raw videos, identifying explicit transitions and indoor–outdoor discontinuities. A curated subset of seamless indoor–outdoor traversals was intentionally preserved to evaluate reconstruction robustness under drastic contextual and illumination shifts. Second, we applied TransNetV2 [24], a state-of-the-art shot boundary detection model, to identify subtle algorithmic or technical cuts that may escape human observation. Finally,

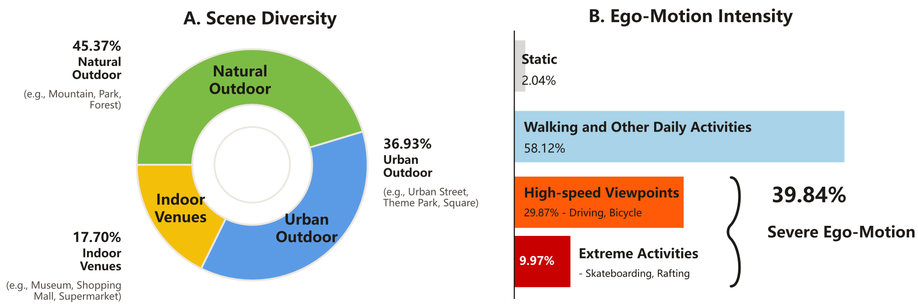

Fig. 3: Overview of Scene and Ego-Motion Diversity in OmniLife360. The percentages indicate the proportion of sequences in each category. (A) Scenes are grouped into three semantic categories: Natural Outdoor, Urban Outdoor, and Indoor Venues. Seamless indoor–outdoor traversals are assigned to the corresponding semantic scene group rather than treated as a separate category. (B)Actions are grouped into four ego-motion types: Static with minimal motion, Walking and Other Daily Activities with mild changes, High-speed Viewpoints with fast translation, and Extreme Activities with abrupt or large motion.

expert auditing was performed to verify all detected boundaries, refine imprecise transitions, and merge automated predictions with manual annotations.

The finalized dataset comprises 2,407 geometrically consistent sequences, spanning approximately 66 hours in total. Unlike prior panoramic benchmarks that focus on short clips or controlled trajectories, OmniLife360 emphasizes minute-level temporal continuity. The average sequence duration reaches 98 seconds, with more than 52% of sequences exceeding two minutes. Such extended trajectories substantially amplify long-term pose uncertainty, cumulative drift, and dynamic-static confusion, thereby enabling principled evaluation of reconstruction robustness under realistic, user-generated capture conditions. Furthermore, all scenes have been processed by COLMAP [22], providing initial sparse point cloud and camera parameters.

3.3 Task-Agnostic Semantic Annotation

To facilitate comprehensive analysis while avoiding interference with geometric supervision, we annotate semantic metadata independently of reconstruction evaluation. Each valid sequence is labeled with fine-grained scene and action categories. To complement discrete tags, we employ state-of-the-art multi-modal large language models (MLLMs) to generate descriptive captions summarizing environmental context, dominant activities, and dynamic characteristics. These captions provide additional interpretability for downstream analysis without affecting reconstruction metrics.

To capture the dataset's characteristics, we roughly categorized and summarized the scene and action distributions (Fig. 3). The OmniLife360 benchmark exhibits significant semantic diversity and ego-motion intensity. It includes 64

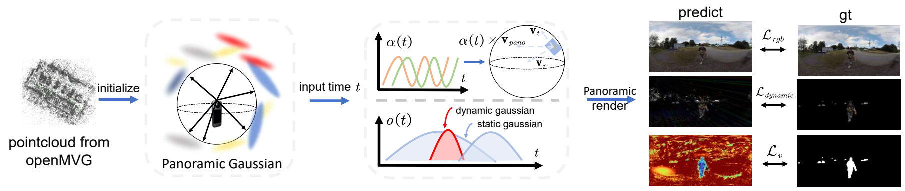

Fig. 4: Pipeline overview. We propose Omni4DGS to optimize the reconstruction performance of panoramic Gaussians in dynamic environments. Given an arbitrary time t as input, we model the opacity $o(t)$ and Gaussian center $\mu(t) = \mu + \alpha(t) \times \boldsymbol{v}_{pano}$ as time-varying values to represent the displacement of dynamic objects as well as their appearance and disappearance along the temporal dimension. To address the distortion problem in the panoramic projection model, we decompose the velocity into radial and tangential velocities for separate learning, and design a latitude-based adaptive weight synthesis to obtain the velocity $\boldsymbol{v}_{pano}$. Furthermore, we achieve explicit supervision for dynamic objects by extracting dynamic object masks offline.

scene categories, spanning urban outdoor, natural outdoor, indoor venues, along with 31 action classes. Static scenes make up a small portion, while walking and everyday locomotion forms the majority. Over 39% of sequences involve intense dynamics, including high-speed viewpoint motion and extreme activities. These dynamics introduce motion blur, rapid camera rotations, and strong ego-motion, providing a challenging testbed for reconstruction robustness. Although the category distribution is not artificially balanced, it reflects real-world usage patterns of consumer panoramic cameras. Outdoor experiences and everyday locomotion naturally dominate user-generated recordings, and preserving this distribution ensures ecological validity rather than enforcing artificial category balancing.

4 Our method: Omni4DGS

Through experiments on our constructed **OmniLife360** dataset, we observe that existing omnidirectional Gaussian-based methods, including feed-forward and per-scene optimization approaches, fail to accurately render dynamic objects in dynamic scenes. To address this limitation, we propose **Omni4DGS**, the first 4D Gaussian Splatting framework tailored for 360° panoramic video input, enabling precise dynamic object reconstruction and artifact-free rendering. Sec. 4.1 introduces the proposed temporal motion formulation for modeling the time-varying Gaussian states. Sec. 4.2 presents dynamic object optimization with 2D mask supervision. Sec. 4.3 describes the training losses, which mitigate rendering artifacts caused by panoramic projection distortion and achieve accurate separation between static background and dynamic foreground objects. The overall pipeline is illustrated in Fig. 4.

4.1 4D Gaussian Splatting in Panoramic Dynamic Scenes

Given video frames captured by a single panoramic camera as input, our goal is to enable accurate rendering at arbitrary viewpoints and time steps. Recent panoramic Gaussian splatting methods typically assume that the scene is static. For example, in ODGS [16], each Gaussian is parameterized by its mean μ, opacity o, color c, the diagonal elements $s \in \mathbb{R}^3$ of the scale matrix, and a rotation represented by a quaternion q. While the formulation is effective for static scene reconstruction, it inherently assumes time-invariant Gaussian representations, therefore cannot model the temporal evolution of dynamic objects commonly present in real-world environments. To address this limitation, we extend the Gaussian parameterization to the temporal domain by introducing periodic harmonic displacement and time-dependent opacity decay [3], allowing Gaussians to evolve smoothly over time and enabling dynamic scene reconstruction.

$$\mu\left(t\right) = \boldsymbol{\mu} + \alpha\left(t\right) \cdot \boldsymbol{v}_{pano} \tag{1}$$

$$o\left(t\right) = o \cdot e^{-\frac{1}{2}\left(\frac{t-\tau}{\beta}\right)^2} \tag{2}$$

Here, $\alpha\left(t\right) = \frac{l}{2\pi} \cdot \sin\left(\frac{2\pi(t-\tau)}{l}\right)$ and τ denotes the peak time of opacity, corresponding to the equilibrium position in periodic motion. β is the lifespan parameter controlling the temporal decay range and duration. $\boldsymbol{v}_{pano} = \left.\frac{d\hat{\boldsymbol{\mu}}(t)}{dt}\right|_{t=\tau} \in \mathbb{R}^3$ controls the motion direction and represents instantaneous velocity at time τ. These parameters are learnable, while l is predefined hyperparameter. This 4D representation enables Gaussians to evolve over time and model dynamic objects.

However, directly using the single learnable velocity vector in Eq. (1) to model Gaussian motion under panoramic cameras leads to inaccurate motion representation. This is because panoramic imaging is based on unit sphere projection, where motion variation better follows angular changes in spherical space rather than linear translation in Euclidean coordinates. A single Euclidean velocity vector cannot explicitly distinguish the geometric differences between radial and tangential motion in the imaging process. As a result, it is difficult to learn motion patterns that conform to the geometric characteristics of panoramic space. To address this issue, we decompose the original velocity $\mathbf{v}_{pano}$ into two independently learnable components: radial velocity $\mathbf{v}_r$ and tangential velocity $\mathbf{v}_t$. We further adopt a latitude-aware adaptive weighting strategy to combine them:

$$\mathbf{v}_{pano} = w_r \mathbf{v}_r + w_t \mathbf{v}_t \tag{3}$$

where $w_t = \left(\sqrt{1 - \hat{\mu}_y^2}\right)^\gamma$ and $w_r = 1 - w_t$. $\hat{\mu}_y$ denotes the normalized y-component of the Gaussian center. On the unit sphere, the latitude angle θ satisfies $\cos\theta = \hat{\mu}_y$. The parameter γ is introduced to enhance the nonlinear distribution of latitude weights, which makes the tangential weights increase more smoothly in the equatorial region and decay more rapidly in the polar regions. This method can explicitly decouple different motion patterns under spherical geometry and adaptively adjust their influences according to latitude, thereby improving the accuracy and stability of dynamic modeling.

4.2 Dynamic Decomposition

A staticness coefficient $\rho = \frac{\beta}{l}$ is used to characterize the dynamic behavior of Gaussians and to separate the static and dynamic parts of the scene in a self-supervised manner. When $\rho \to 0$, opacity exhibits short-term oscillations and reaches a maximum at a time close to τ to indicate dynamic objects. However, we found that under panoramic images, accurate static-dynamic decoupling cannot be achieved solely through the ℓ_1 self-supervised loss with the input RGB images, which leads to confusion between static and dynamic components and results in rendering artifacts. To address this, we use OmniSAM [31], a semantic segmentation model specifically designed for panoramic images, to extract dynamic objects (*e.g.*, pedestrians, cars) from the image as the motion mask prior M_d. We use the dynamic regions in the image as supervision for the rendering results of dynamic Gaussians, ensuring that dynamic Gaussians only model the corresponding dynamic objects. Specifically, Gaussians with $\rho < \theta_{d_{gm}}$ are regarded as dynamic Gaussians $\mathcal{G}_d$. Following the rendering pipeline proposed in ODGS, we obtain the rendering results $\hat{C}_d$ corresponding to the dynamic Gaussians, and use ℓ_1 loss and SSIM [26] to supervise the dynamic objects.

$$\mathcal{L}_d = (1-\lambda_d)\cdot\mathcal{L}_1\left(\hat{C}_d,\ M_d\cdot C_{gt}\right) + \lambda_d\cdot SSIM\left(\hat{C}_d,\ M_d\cdot C_{gt}\right) \tag{4}$$

where $M_d \cdot C_{gt}$ represents the pixel values of the dynamic regions in the original image. To further decouple the dynamic and static parts, we modify the ODGS [16] rasterizer operator to support feature rendering, such as opacity map, velocity map, and depth map:

$$F = \sum_{i\in\mathcal{N}} f_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \tag{5}$$

where f_i denotes any Gaussian-associated feature (*e.g.*, opacity, depth, velocity), and $\mathcal{N}$ is the number of Gaussians. We use ℓ_1 loss to make the opacity map obtained from dynamic Gaussians consistent with M_d:

$$\mathcal{L}_o = \|O_d - M_d\|_1 \tag{6}$$

The total loss for dynamic Gaussians is:

$$\mathcal{L}_{dynamic} = \mathcal{L}_d + \mathcal{L}_o \tag{7}$$

The additional supervision on dynamic Gaussians effectively alleviates the inaccurate static-dynamic decoupling problem that occurs when dynamic Gaussians are learned solely from RGB self-supervision in panoramic images.

4.3 Training Optimization

The $\mathcal{L}_{dynamic}$ used in Sec. 4.2 ensures supervision for dynamic Gaussians. However, the static background occupies most of the area in most images, so we follow vanilla 3DGS [13] and use conventional rendering losses to obtain a stable static background:

$$\mathcal{L}_{rgb} = (1 - \lambda_{rgb}) \cdot \mathcal{L}_1(\hat{C}, C_{gt}) + \lambda_{rgb} \cdot SSIM(\hat{C}, C_{gt}) \tag{8}$$

Velocity Regularization Given the velocity $\mathbf{v}_{pano}$, we follow [3] to obtain the average velocity $\hat{\mathbf{v}}_{pano} = \mathbf{v}_{pano} \cdot \exp\left(-\frac{\rho}{2}\right)$. To further constrain the velocity in static regions to be close to 0, we use the formula Eq. (5) to obtain the average velocity map $\hat{V}$, and use the following loss as a regularization term:

$$\mathcal{L}_v = \left\| \hat{V} \cdot (1 - M_d) \right\|_1 \tag{9}$$

Motion Regularization We observe that under panoramic imaging, small angular perturbations of Gaussian centers in 3D space along the spherical direction will be nonlinearly amplified in ERP projection, leading to significant pixel shifts and further exacerbating static-dynamic decoupling difficulties. We further constrain the displacement of static Gaussians. For Gaussians with $\rho \geq \theta_{d_{gm}}$, we can obtain the normalized direction of the center coordinate at time t:

$$d(t) = \frac{\mu_s(t)}{\|\mu_s(t)\|} \tag{10}$$

We impose constraints on the direction changes between adjacent time steps:

$$\mathcal{L}_m = \|d(t + \Delta t) - d(t)\|_2^2 \tag{11}$$

The total loss used for training is:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{dyn}\mathcal{L}_{dynamic} + \lambda_v\mathcal{L}_v + \lambda_m\mathcal{L}_m \tag{12}$$

5 Experiments

5.1 Experimental Settings

Dataset We perform extensive evaluations on our OmniLife360 benchmark, selecting 111 representative scenes that include various scene types. To accommodate hardware constraints and prevent Out-Of-Memory (OOM) issues across different baselines, we downsample the video frames to 1K–2K resolution. For sequences without calibrated trajectories, we obtain camera poses and sparse point clouds using a unified COLMAP-based SfM pipeline. We apply the same pose-validation procedure to all evaluated scenes, and reprocess unstable reconstructions with a more robust configuration before inclusion.

Table 2: Quantitative comparisons for per-scene optimization methods on OmniLife360. We report metrics on both the Input View set (reconstruction fidelity) and the Novel View set (novel view synthesis). The best results are highlighted in **bold**.

Method	Input View			Novel View		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ODGS [16]	23.727	0.792	0.385	20.145	**0.712**	0.433
Ours	**29.655**	**0.896**	**0.146**	**20.640**	0.624	**0.335**

Table 3: Quantitative comparisons for generalizable feed-forward methods on OmniLife360. We test four feed-forward methods in the selected set. The best results are highlighted in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Splatter-360 [5]	14.081	0.395	0.555
PanSplat [28]	15.774	0.526	0.493
PanoSplatt3R [21]	**17.646**	0.521	**0.450**
OmniSplat [17]	17.270	**0.543**	0.472

Comparing Methods We compare our approach with two main paradigms. For per-scene optimization baselines, we select ODGS [16]. For generalizable feed-forward methods, we evaluate recent state-of-the-art architectures including PanSplat [28], OmniSplat [17], PanoSplatt3R [21], and Splatter-360 [5].

Metrics Following standard novel view synthesis protocols, we quantitatively assess rendering quality using three standard metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [26] to assess low-level signal and structural fidelity, and Learned Perceptual Image Patch Similarity (LPIPS) [29] to measure high-level perceptual quality.

5.2 Implementation Details

All 111 scenes go through COLMAP process to generate initial point cloud and camera parameters. We also perform corresponding transformations to meet the need for some methods that require the form of OpenMVG [19] or HM3D [20].

Although videos are predominantly 4K or 8K, we resize and pad them to unified dimensions to maintain consistent aspect ratios. For per-scene optimization methods (including ours), we train for 30K iterations using a 0.8/0.2 train-test split for each sequence. For feed-forward methods, we adopt a sparse-view interpolation protocol: we uniformly sample consecutive frames into groups of 10, using the first and last frames as input to predict the intermediate 8 novel views. Since optimization-based methods (dense reconstruction) and feed-forward methods (sparse few-shot inference) are evaluated under distinct protocols, their absolute scores are not strictly cross-comparable. However, the relative

Table 4: Quantitative results on four representative dynamic sequences.

Methods	Metrics	Avg.	Market	Lake	Mountain	Urban Street
PanSplat	PSNR↑	15.988	13.183	17.136	17.923	15.711
	SSIM↑	0.553	0.401	0.661	0.583	0.566
	LPIPS↓	0.469	0.567	0.412	0.413	0.485
Splatter-360	PSNR↑	15.017	12.660	15.347	17.439	14.621
	SSIM↑	0.443	0.318	0.508	0.492	0.455
	LPIPS↓	0.517	0.595	0.477	0.460	0.536
OmniSplat	PSNR↑	17.356	14.354	19.000	19.391	16.677
	SSIM↑	0.572	0.402	0.688	0.607	0.590
	LPIPS↓	0.454	0.568	0.379	0.395	0.474
PanoSplatt3R	PSNR↑	17.797	13.507	21.443	20.816	15.421
	SSIM↑	0.545	0.359	0.709	0.592	0.518
	LPIPS↓	0.437	0.573	0.318	0.361	0.496
ODGS	PSNR↑	20.492	16.789	22.596	22.048	20.535
	SSIM↑	**0.733**	**0.624**	**0.792**	**0.748**	**0.768**
	LPIPS↓	0.416	0.505	0.333	0.404	0.422
Ours	PSNR↑	**21.519**	**17.849**	**23.727**	**22.776**	**21.722**
	SSIM↑	0.674	0.540	0.766	0.700	0.691
	LPIPS↓	**0.302**	**0.381**	**0.248**	**0.264**	**0.313**

performance within each paradigm clearly demonstrates algorithmic superiority. All experiments run on NVIDIA RTX 4090 GPUs.

5.3 Experiment Results

Results on OmniLife360 testset Quantitative results are summarized in Tab. 2 and Tab. 3. We use ODGS as the primary panoramic optimization baseline, and further adapt representative dynamic 4DGS methods to panoramic projection for additional comparison. Our proposed Omni4DGS demonstrates a significant improvement over ODGS. Specifically, for reconstruction fidelity (Input View), our method achieves a remarkable boost in PSNR from **23.727 dB to 29.655 dB**, alongside a drastic reduction in LPIPS from **0.385 to 0.146**. This strong performance extends to novel view synthesis, where Omni4DGS consistently outperforms ODGS in PSNR **(20.640 vs. 20.145)** and exhibits a noticeably lower LPIPS **(0.335 vs. 0.433)**. Compared to the best-performing feed-forward method, PanoSplatt3R (PSNR of 17.646 and LPIPS of 0.450), our per-scene optimization framework maintains a commanding margin in both PSNR and LPIPS, underscoring its superiority for high-quality novel view synthesis.

Results on four typical dynamic scenes We select four representative dynamic scenes, namely *Market*, *Lake*, *Mountain*, and *Urban Street*. Each scene

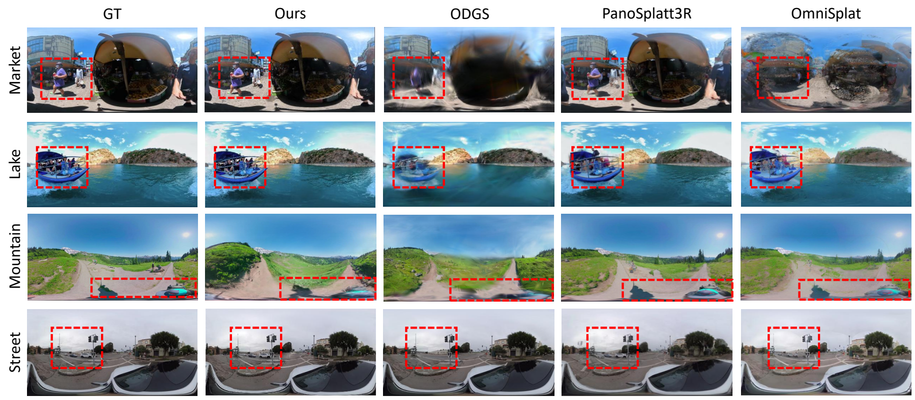

Fig. 5: Rendering comparisons in four typical dynamic scenes. Our method introduces the temporal dimension and applies supervision on dynamic objects to achieve accurate reconstruction.

contains multiple sequences captured under the same environment. The quantitative results for the four scenes are reported in Tab. 4, where the values for each scene represent the average results over multiple sequences within that scene. Our method achieves the best performance across all scenes. For qualitative comparison, we select one representative sequence from each scene, and the rendering results are shown in Fig. 5. ODGS produces severe blurring when rendering dynamic objects. The other two feed-forward methods often exhibit ghosting artifacts. In contrast, our method introduces a temporal dimension and explicitly models dynamic objects, enabling artifact-free rendering results.

Comparison with Existing 4DGS Methods To the best of our knowledge, no existing 4D Gaussian Splatting method is designed for dynamic reconstruction from panoramic imagery. We therefore compare our method with two representative pinhole-camera-based 4DGS methods, Deformable-GS [27] and PVG [3], by replacing their projection models with panoramic projection. We evaluate these methods on four representative dynamic panoramic scenes. Quantitative results in Tab. 5 show that our method consistently outperforms existing 4DGS approaches, with additional qualitative comparisons provided in the Appendix. Existing methods fail to model panoramic distortions, causing inaccurate motion estimation and rendering artifacts, whereas our panoramic-aware motion modeling effectively addresses these issues.

Table 5: Comparison with existing 4DGS methods.

Method	Metric	*bike*	*community*	road	*square*
ODGS	PSNR↑	10.137	16.641	15.618	19.510
Deformable-GS*	PSNR↑	11.183	15.154	15.064	17.063
PVG*	PSNR↑	12.379	17.168	17.484	21.020
Ours	PSNR↑	**12.930**	**18.063**	**17.533**	**21.941**

Table 6: Ablation study of Omni4DGS. We report the average performance over four dynamic panoramic scenes.

Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**
w/o temporal opacity	19.865	0.631	0.324
w/o temporal motion	20.999	0.666	0.299
w/o velocity decom.	21.191	0.664	0.300
w/o dynamic mask	20.753	0.651	0.302
w/o motion reg.	21.023	0.659	0.302
Ours (full)	**21.473**	**0.679**	**0.285**

5.4 Ablation Study

We conduct ablation studies on four representative dynamic panoramic scenes, including *Space*, *Lake*, *Market*, and *Urban Street*. As shown in Tab. 6, the full model achieves the best average performance across PSNR, SSIM, and LPIPS, demonstrating the effectiveness of the proposed components.

Removing temporal opacity or temporal motion leads to clear performance degradation, confirming the importance of modeling time-varying visibility and Gaussian displacement for dynamic panoramic reconstruction. The drop caused by removing velocity decomposition further validates the benefit of spherical-aware motion modeling under ERP projection. In addition, removing dynamic mask supervision weakens the separation between dynamic foreground and static background, while removing motion regularization reduces the stability of background Gaussians. Overall, these results demonstrate that the proposed components provide complementary gains for robust dynamic 360° reconstruction.

6 Conclusion

In this paper, we present **OmniLife360**, a large-scale benchmark for 3D reconstruction from in-the-wild 360° captures that systematically integrates long-horizon temporal continuity, diverse real-world dynamics, and strict geometric validity. Through comprehensive evaluation, we reveal critical limitations of existing panoramic reconstruction methods under strong ego-motion and pervasive dynamic interference. To address these challenges, we propose **Omni4DGS**, a spherical-aware 4D Gaussian Splatting framework with motion decomposition and mask-supervised dynamic-static decoupling, enabling more stable and artifact-free reconstruction in complex panoramic scenes. We hope OmniLife360 will foster future research toward robust, dynamic-aware, and scalable 360° scene reconstruction in realistic deployment scenarios.

References

1. Bai, J., Huang, L., Guo, J., Gong, W., Li, Y., Guo, Y.: 360-gs: Layout-guided panoramic gaussian splatting for indoor roaming. arXiv preprint arXiv:2402.00763 (2024)
2. Bertel, T., Yuan, M., Lindroos, R., Richardt, C.: Omniphotos: casual 360 vr photography. ACM Transactions on Graphics (TOG) **39**(6), 1–12 (2020)
3. Chen, Y., Gu, C., Jiang, J., Zhu, X., Zhang, L.: Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. International Journal of Computer Vision (2026)
4. Chen, Z., Cao, Y.P., Guo, Y.C., Wang, C., Shan, Y., Zhang, S.H.: Panogrf: Generalizable spherical radiance fields for wide-baseline panoramas. arXiv preprint arXiv:2306.01531 (2023)
5. Chen, Z., Wu, C., Shen, Z., Zhao, C., Ye, W., Feng, H., Ding, E., Zhang, S.H.: Splatter-360: Generalizable 360 gaussian splatting for wide-baseline panoramic images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21590–21599 (June 2025)
6. Choi, C., Kim, S.M., Kim, Y.M.: Balanced spherical grid for egocentric view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16590–16599 (2023)
7. Choi, D., Jang, H., Kim, M.H.: Omnilocalrf: Omnidirectional local radiance fields from dynamic videos. In: CVPR (June 2024)
8. Huang, H., Chen, Y., Li, L., Cheng, H., Braud, T., Zhao, Y., Yeung, S.K.: Sc-omnigs: Self-calibrating omnidirectional gaussian splatting. In: The Thirteenth International Conference on Learning Representations (2025)
9. Huang, H., Chen, Y., Zhang, T., Yeung, S.K.: 360roam: Real-time indoor roaming using geometry-aware 360° radiance fields. arXiv preprint arXiv:2208.02705 (2022)
10. Huang, H., Liu, C., Zhu, Y., Hui, C., Braud, T., Yeung, S.K.: 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
11. Huang, H., Yeung, S.K.: 360vo: Visual odometry using a single 360 camera. In: 2022 international conference on robotics and automation (icra). pp. 5594–5600. IEEE (2022)
12. Kageyama, A., Sato, T.: "yin-yang grid": An overset grid in spherical geometry. Geochemistry, Geophysics, Geosystems **5**(9) (2004)
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 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/`, accessed: 2026-06-30
14. Kim, J., Choi, C., Jang, H., Kim, Y.M.: Piccolo: Point cloud-centric omnidirectional localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3313–3323 (2021)
15. Kulkarni, S., Yin, P., Scherer, S.: 360fusionnerf: Panoramic neural radiance fields with joint guidance. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7202–7209 (2023). `https://doi.org/10.1109/IROS55552.2023.10341346`
16. Lee, S., Chung, J., Huh, J., Lee, K.M.: Odgs: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings. Advances in Neural Information Processing Systems (NeurIPS) **37** (2024)

17. Lee, S., Chung, J., Kim, K., Huh, J., Lee, G., Lee, M., Lee, K.M.: Omnisplat: Taming feed-forward 3d gaussian splatting for omnidirectional images with editable capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16356–16365 (June 2025)
18. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. arXiv preprint arXiv:2109.13410 (2021)
19. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: Openmvg: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 60–74. Springer (2016)
20. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
21. Ren, J., Xiang, M., Zhu, J., Dai, Y.: Panosplatt3r: Leveraging perspective pretraining for generalized unposed wide-baseline panorama reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28959–28969 (October 2025)
22. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016). https://doi.org/10.1109/CVPR.2016.445
23. Shin, C., Cho, W.O., Kim, S.J.: Seam360gs: Seamless 360deg gaussian splatting from real-world omnidirectional images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28970–28979 (2025)
24. Souček, T., Lokoč, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. arXiv preprint arXiv:2008.04838 (2020)
25. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
26. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
27. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023)
28. Zhang, C., Xu, H., Wu, Q., Gambardella, C.C., Phung, D., Cai, J.: Pansplat: 4k panorama synthesis with feed-forward gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11437–11447 (June 2025)
29. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
30. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Proceedings of The European Conference on Computer Vision (ECCV) (2020)
31. Zhong, D., Zheng, X., Liao, C., Lyu, Y., Chen, J., Wu, S., Zhang, L., Hu, X.: Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23892–23901 (2025)