This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

WANDERLAND: Geometrically Grounded Simulation for Open-World Embodied AI

Xinhao Liu[*,1] Jiaqi Li[*,1] Youming Deng[2] Ruxin Chen[1] Yingjia Zhang[1]
Yifei Ma[1] Li Guo[1] Yiming Li[1] Jing Zhang[✉,1] Chen Feng[✉,1]
New York University[1] Cornell University[2]

https://ai4ce.github.io/wanderland/

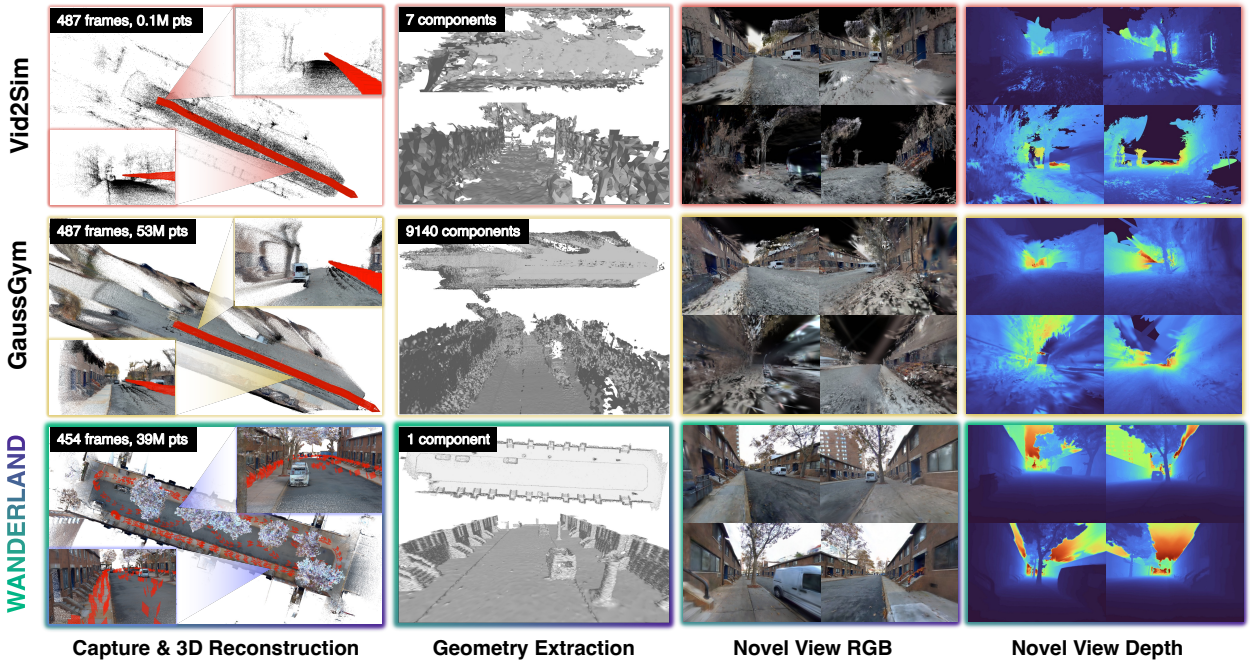

Figure 1. **Do video-3DGS frameworks provide geometrically grounded and photorealistic simulation?** We demonstrate that building such simulations from casually captured touring videos often fails due to limited view diversity, inaccurate 3D reconstruction, unreliable geometry extraction, and degraded novel-view extrapolation. We propose the WANDERLAND framework that features multi-sensor diverse-view capture, reliable reconstruction, accurate metric-scale geometry, and robust view synthesis.

Abstract

Reproducible closed-loop evaluation remains a major bottleneck in Embodied AI such as visual navigation. A promising path forward is high-fidelity simulation that combines photorealistic sensor rendering with geometrically grounded interaction in complex, open-world urban environments. Although recent video-3DGS methods ease open-world scene capturing, they are still unsuitable for benchmarking due to large visual and geometric sim-to-real gaps. To address these challenges, we introduce Wanderland, a real-to-sim framework that features multi-sensor capture, reliable reconstruction, accurate geometry, and robust view synthesis. Using this pipeline, we curate a diverse dataset of indoor-outdoor urban scenes and systematically demonstrate how image-only pipelines scale poorly, how geometry quality impacts novel view synthesis, and how all of these adversely affect navigation policy learning and evaluation reliability. Beyond serving as a trusted testbed for embodied navigation, Wanderland's rich raw sensor data further allows benchmarking of 3D reconstruction and novel view synthesis models. Our work establishes a new foundation for reproducible research in open-world embodied AI.

[*]Equal contribution, random order.
✉ Corresponding authors, {jz6676,cfeng}@nyu.edu

1. Introduction

Embodied AI research has advanced significantly in recent years, driven by high-performance simulation platforms [1–6] and datasets [7–12]. They have played a central role in this progress by standardizing tasks, metrics, and providing closed-loop evaluation. With the emergence of foundation models for vision, language, and decision making [13–16], it is natural to extend embodied AI beyond domestic interiors to open-world settings. Applications in embodied navigation include last-mile delivery [17], campus-scale wayfinding [18–20], and service robots that must traverse lobbies, hallways, plazas, and sidewalks [21–23]. These use cases require simulative environments with large spatial extents, mixed indoor-outdoor coverage, high-fidelity sensor simulation, and reliable physics for interaction. They also require datasets with diverse scene coverage, such as sidewalks, grocery stores, or subway stations. *How can we build a high-fidelity real-to-sim testbed to evaluate and benchmark open-world navigation systems?*

This demand poses significant challenges to classic datasets and their curation pipelines using RGB-D sensors. These challenges begin at capture: prevalent RGB-D sensors fail outdoors due to sunlight interference and limited range, and tripod-based methods are inefficient for large-scale areas. During reconstruction, pose estimation via RGB-D fusion is prone to drift and failure in large-scale, low-texture, and less-structured environments. Finally, simulation fidelity is further affected by low-quality texture meshes reconstructed without enough spatial resolution, or with artifacts like non-watertight and fragmented geometry, especially outdoors. Consequently, these curation pipelines cannot reliably support large-scale, high-fidelity simulations in open-world environments.

To overcome the limitations of indoor-only datasets, recent work has explored using online videos as an alternative data source for building open-world environments. Methods based on 3D Gaussian Splatting (3DGS) [31] have shown promise in creating visually compelling simulations from casual videos and enabling navigation policy training via reinforcement learning [32–35]. However, these video-based approaches face fundamental challenges that prevent their use as standardized benchmarks:

First, they suffer from **inaccurate 3D reconstruction** due to reliance on pure RGB-based methods like SfM [28, 36] or deep reconstruction models [30, 37], yielding non-metric camera poses and depth estimations. Second, they produce **unreliable geometry** for physical interaction simulation, with collision meshes extracted from 3DGS opacity fields being fragmented and metrically ungrounded [38, 39]. Finally, they exhibit severe **extrapolated view degradation** due to uniform video trajectories, causing rendering quality to sharply decline for viewpoints beyond the capture path [40]. Consequently, while suitable for training, these environments lack the geometric grounding required for reproducible benchmarking of embodied navigation systems.

To bridge this gap demonstrated in Fig. 1, we present WANDERLAND, a robust real-to-sim framework featuring multi-sensor diverse-view capture, reliable reconstruction, accurate metric-scale geometry, and robust view synthesis. Our pipeline begins with data capturing with a handheld multi-sensor 3D scanner (Fig. 2). This setup enables dense, multi-view capture across large-scale indoor and outdoor scenes, providing rich and diverse visual coverage. Our reconstruction leverages a LiDAR-inertial-visual (LIV) SLAM system, which fuses these complementary sensor streams to produce globally consistent, metric-scale point clouds and highly accurate camera poses.

This reliable geometric foundation directly addresses the limitations of video-only pipelines, ensuring metric accuracy and completeness. We then utilize this precise geometry to initialize high-quality 3DGS models and extract clean, reliable collision meshes. Finally, the framework integrates the 3DGS model and geometrically grounded mesh into Isaac Sim, creating a unified environment that supports both photorealistic rendering and geometrically grounded interaction. As shown in Tab. 1, our approach achieves the best of both worlds: efficient open-world capture and geometrically accurate reconstruction.

Leveraging this framework, we introduce the WANDERLAND dataset, which serves not only as a geometrically grounded and photorealistic simulation environment, but also as a critical diagnostic tool for understanding the limitations of existing approaches. We use the dataset to systematically demonstrate how and why vision-only reconstruction methods fail to provide reliable geometric groundings for embodied AI, particularly in open-world settings where metric accuracy and consistent novel view synthesis are essential.

Beyond identifying these issues, WANDERLAND establishes a trusted testbed for benchmarking embodied navigation tasks like image-goal and vision-language navigation under closed-loop control, enabling reliable evaluation in environments that are both visually realistic and geometrically consistent. Furthermore, our rich raw sensor data provides a valuable benchmark for core 3D vision tasks, supporting comprehensive evaluation of 3D reconstruction and novel view synthesis methods. This multi-faceted design enables ablation studies on simulation construction and offers key insights towards advancing large-scale 3D vision and embodied AI research.

We summarize our contributions as follows:

1. We identify fundamental limitations of video-3DGS pipelines for embodied AI and introduce a new real-to-sim framework that overcomes these challenges.
2. We introduce WANDERLAND framework and dataset for systematic analysis of key factors in simulation construc-

Table 1. **Real-to-sim datasets for embodied navigation**. Classic datasets are mostly from dedicated tripod-based capture, providing metric scale environments, and representing scenes as textured meshes, but they are *limited to indoor scenes*. 3DGS datasets can include outdoor scenes, but they suffer from *inaccurate 3D reconstruction* and *non-metric-scale* environments. *: Most scenes are from ARKitScenes [24] and ANYmal GrandTour datasets [25]. : Tripod-based capture. : Casual online videos. : Generated videos. : Mobile 3D scanner.

Dataset	Source	Capture	Scene	Metric Scale	#Scenes	#Frames	Scene Size (m^2)	Geometry Source	Reconstruction
Classic Datasets									
Matterport3D [9]		RGB-D	Indoor	✓	90	11K	102K	Depth	Matterport
ScanNet [11]		RGB-D, IMU	Indoor	✓	1,513	2.5M	40K	Depth	BundleFusion [26]
Gibson [12]		RGB-D	Indoor	✓	572	—	218K	Depth	Proprietary
HM3D [7]		RGB-D	Indoor	✓	1,000	—	365K	Depth	Matterport
3DGS Datasets									
Vid2Sim [27]		RGB	Outdoor	×	30	13.5K	—	3DGS Opacity	GLOMAP [28]
GaussGym [29]		RGB	Mixed	×	2,500*	—	—	Point Map	VGGT [30]
WANDERLAND (ours)		**LiDAR, IMU, RGB**	**Mixed**	✓	**530**	**420K**	**3.8M**	**LiDAR**	**LIV SLAM**

tion. It also serves as a testbed for embodied navigation.

3. We highlight that visual-only reconstructions remain significantly less accurate than LiDAR-enhanced ones, resulting in unreliable simulation for embodied AI.
4. We provide rich raw sensor data that enables benchmark evaluation for 3D reconstruction and novel view synthesis, supporting method evaluation and ablation studies.

2. Related Work

Real-to-sim datasets for embodied navigation. Classic embodied navigation datasets such as Matterport3D [9], ScanNet [11], and HM3D [7] provide metric-scale environments through dedicated RGB-D capture and mesh-based reconstruction, but remain limited to indoor settings. In contrast, our WANDERLAND dataset employs handheld multi-sensor capture and 3DGS representation, enabling high-fidelity sensor simulation across diverse indoor-outdoor environments while maintaining metric accuracy.

Recent video-based approaches expand to open-world settings using alternative data sources. Vid2Sim [27] reconstructs 30 outdoor scenes from 13.5K YouTube frames using GLOMAP for SfM initialization and 3DGS opacity for geometry. GaussGym [29] aggregates 2,500 mixed scenes from online and generated videos, employing VGGT [30] and NKSR [41] for reconstruction. Compared to these vision-only methods that rely on visual inference from casual video sources, WANDERLAND utilizes multi-modal sensing and LIV-SLAM reconstruction, providing geometrically accurate foundations for physical interaction while achieving visual fidelity through 3DGS rendering.

Open-world 3D reconstruction. Robust 3D reconstruction for open-world environments remains challenging due to the limitations of existing capture methodologies. Many outdoor scene datasets [42, 43] rely on SfM methods like COLMAP [36] for sparse reconstruction. Due to its vision-only nature, SfM methods can't provide metric scale camera poses and are prone to errors. In contrast to the image set input in SfM, SLAM algorithms take sequential input. Visual(-inertial) SLAM systems [44–49] improve tracking but still lack absolute scale and global consistency over long trajectories. LiDAR(-inertial) SLAM [50–54] provides metric accuracy and geometric precision, but lacks the semantics from RGB input. Our work leverages LIV-SLAM [55–57] to enable efficient, handheld capture of large-scale environments while overcoming the individual limitations of each modality.

Photorealistic sensor simulation. Traditional approaches rely on textured mesh for sensor simulation [7, 9, 11, 12, 24, 58]. While effective for indoor assets, meshes miss fine structures, require aggressive decimation to scale, and cannot encode view-dependent appearance, which limits photorealistic rendering in unbounded scenes. Neural representations like NeRF [32, 59] and 3DGS [27, 33, 34] enable high-quality rendering. However, image-only training introduces scale ambiguity and unstable geometry in neural representations. Surfaces extracted from densities or opacities [38, 39] are often noisy or incomplete, weakening collision reliability. Our pipeline combines both worlds: LiDAR provides metric-scale geometry for collision meshes and 3DGS initialization, while 3DGS trained with multi-view images with LIV-SLAM poses enable high-fidelity rendering. This hybrid representation ensures geometrically grounded interaction with photorealistic sensor simulation.

Embodied visual navigation. A significant body of work in navigation focuses on goal-conditioned navigation [60–62]. A variety of methods have been developed, ranging from end-to-end Reinforcement Learning [63] to more structured, modular approaches built upon explicit spatial memories [64–66]. Real-world embodied navigation extends these ideas to deploy on quadrupeds or humanoids [18, 21, 22], but evaluation is often limited to fixed routes or logs due to collection cost, safety, and a lack of open-world, simulator-ready assets. Vision-language navigation (VLN) frames the task as following natural-language instructions. Datasets built from crowd-sourced descriptions [67–69] have advanced this line, yet instruction quality, alignment to visual evidence, and indoor-only layouts are still main constraints. VLN policies range from early se-

(a) Mobile app. (b) Device specifications.

Figure 2. **The MetaCam device**. (a) SkylandX MetaCam Air used for data collection, equipped with a companion app for mobile capture. (b) Working frequency of each sensor.

quence models [70] to latest VLA models [71]. Our dataset supports both families. We generate optimal expert trajectories by computing the shortest path on the navigation mesh. We then leverage MLLM to automatically annotate these traversal videos with rich, contextual descriptions.

3. WANDERLAND

3.1. Data Collection

The MetaCam device. We collect data using MetaCam[1], a compact commercial handheld 3D scanner as illustrated in Fig. 2. The device integrates a suite of factory-calibrated sensors: a Livox Mid-360 non-repetitive LiDAR (with a built-in IMU), an RTK-GNSS antenna, and two synchronized 4K fisheye cameras with over 180° field of view. The LiDAR is mounted at a tuned inclination to optimize the capture of ground-level details and maximize the field-of-view (FOV) overlap with the cameras. We also utilize the accompanying mobile app that provides a real-time preview of the colorized point cloud on a mobile device, enabling operators to actively fill any gaps in the environment during capture, ensuring complete scene coverage.

Collection protocol. Our data collection scenes are selected to have optimal sizes of 5,000–10,000 square meters to balance complexity and coverage. Instead of recording at a fixed frame rate, we trigger RGB capture whenever the device has moved by a fixed distance or rotated beyond a fixed angular threshold, which leads to more uniformly distributed viewpoints throughout each scene. As shown in Fig. 3, each capture session employs a systematic trajectory strategy: training trajectories follow closed-loop paths with dense, multi-view coverage of all navigable areas, while extrapolation trajectories simulate natural navigation paths with minimal overlap. This differs from city touring videos, such as those used in Vid2Sim [27] and GaussGym [29], where the camera views are mostly uni-directional as shown in Fig. 1 (first column, top two rows). To ensure data quality, we minimize dynamic obstacles and reflective surfaces,

[1] https://skylandx.com/metacam-air/

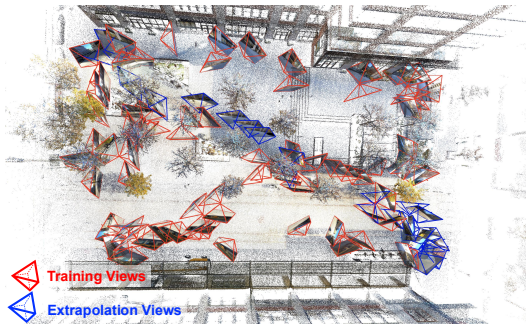

Figure 3. **Data collection trajectory**. To facilitate both diverse-view capture and evaluation of extrapolated views in navigation, our data is collected with well-defined training and extrapolation splits. Training views ensure accurate reconstruction, while extrapolation views are used for evaluation.

maintain consistent lighting conditions during each session, and perform real-time point cloud monitoring to verify coverage completeness.

Time and location. We collected data across diverse indoor and outdoor urban environments in New York City and Jersey City. Our dataset encompasses distinct scenes spanning residential buildings, business districts, public streets, plazas, and university campuses. More details are shown in Appendix Fig. I. To ensure appearance diversity and robustness, data was captured across different times of day (*e.g.,* morning, noon, dawn) and under varying weather conditions (*e.g.,* sunny, overcast, light rain).

3.2. Data Processing

Mapping and Reconstruction. We process the raw sensor data using MetaCam Studio, which implements a robust LiDAR-inertial-visual-GNSS sensor fusion pipeline. While the implementation is proprietary, the methodology builds upon established multi-sensor fusion systems [49, 55, 72], extending them for large-scale urban mapping. This sophisticated fusion framework produces dense, metric-scale point clouds with high accuracy and globally consistent camera trajectories as shown in Fig. 4. This provides the geometric foundation for all subsequent steps.

Image masking. We apply a two-stage masking strategy for our raw fisheye images. First, we use Egoblur [73] to mask out human faces and car plates to ensure privacy and anonymity for data release. Second, we use an object detector [74] to mask out common dynamic objects including people, animals, and vehicles, which are used to filter out invalid pixels during 3DGS training.

Image undistortion. While a large FOV can provide rich coverage of the environment, we crop raw fisheye images into 120° and undistort into perspective views for 3DGS training. This is due to the limitation from the low-order approximation of camera distortion, and is a common practice in related work [75, 76]. It also ensures better masking

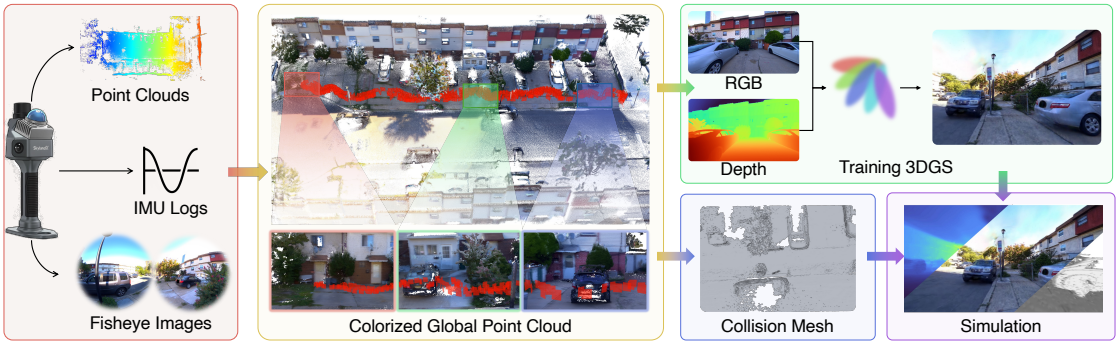

Figure 4. **Data Processing Pipeline**. Our pipeline begins with multi-sensor capture using the MetaCam device in real-world urban spaces. MetaCam Studio processes the raw data via LIV-SLAM to produce a colorized, globally consistent metric point cloud and accurate camera poses. We then initialize 3D Gaussians from the metric point cloud and render per-view depth maps from this initialization. The 3DGS model is optimized with both photometric and depth losses. In parallel, we extract a reliable collision mesh from the same global point cloud. Finally, we integrate the trained 3DGS model and the collision mesh into a single Universal Scene Description (USD) scene, which can be directly loaded into Isaac Sim for training and evaluating navigation policies

results as the models are mostly trained on pinhole images. More details in Appendix Sec. B.

3.3. Training 3D Gaussians

Initialization. We initialize 3D Gaussians from the dense colorized global point cloud produced by MetaCam Studio. Due to its millimeter-level density and accuracy, it provides an high-quality initialization for subsequent 3DGS training. We parameterize Gaussian opacity as inversely proportional to volumetric density obtained by a KNN heuristic. This reduces the dominance of large Gaussians or spurious floaters in initial training steps.

Depth regularization. Many works have shown depth loss helps 3DGS training [27, 38]. In contrast to use monocular depth as pseudo ground truth, we directly project the initialized Gaussians to each camera pose and treat them as ground truth depth. While it is intuitive to freeze the Gaussian center attributes considering the accurate initialization, this rigid parameterization imposes an overly uniform geometric prior that limits the model's ability to capture high-frequency details from images, leading to degraded visual quality. We still follow the common practice of combining photometric and depth losses. As a result, we can stabilize 3DGS training and well align the trained Gaussians to the scene geometry.

Training view augmentation. We also utilize generative models to enhance extrapolated view synthesis. We use the pretrained model from Difix3D+ [77] to augment training views with clean and geometrically accurate novel views. We gradually expand the augmented views away from the training views along the training steps. This is crucial for stabilizing large-scale 3DGS training and improving sensor simulation for extrapolated camera views. More details in Appendix Sec. B.

3.4. Geometrically Grounded Simulation

Mesh extraction. To achieve geometric grounding, we extract mesh from the dense global point cloud. We voxelize the point cloud as an occupancy grid and use the marching cubes [78] algorithm to obtain geometric meshes. To further improve rendering performance, we remove parts that far from the collection trajectory and filter out fragments with small number of faces.

Scene integration. As both the mesh and 3DGS model are based on the global point cloud and share the same coordinate system, we can integrate them into the Unified Scene Description (USD) format. In this representation, the mesh provides a lightweight physics and collision layer, while the 3DGS model is used as the primary renderer. The scene can be directly loaded into Isaac Sim for training and evaluating embodied navigation systems.

3.5. Defining Navigation Tasks

Expert trajectory. We derive navigation trajectories from mesh geometry. We import the mesh into Unity and utilize the NavMesh baking API to extract a triangulated navigable surface. We then use the pathfinding module to generate collision-free expert trajectories based on the navmesh. The starting and goal positions are sampled in the vicinity of capturing cameras. These trajectories can support point-goal and image-goal navigation tasks.

Language instruction. For VLN, we additionally generate navigation instructions for each trajectory. We replay each trajectory in the simulation and generate an egocentric video. We use a VLM [84] to generate natural language instructions based on the video. The instructions are further verified by humans for reliability. Compared to fully manual annotation [67, 68], this procedure is substantially more scalable and yields more stable instructions across scenes, while still allowing for quality control. Detailed steps are

Table 2. **Vision-based reconstruction still underperforms LIV-SLAM**. All methods are evaluated on the same number of images for each scene. T-ATE^{R} (raw) and T-ATE^{S} (scaled) means ATE evaluated *without/with* ground-truth scale alignment. COLMAP^{calib} means COLMAP with ground truth intrinsic calibration. SR stands for success rate. Detailed definition on evaluation metrics is described in Appendix Sec. E.1. For each entry, we show the **mean/median** across the test dataset.

Method	**T-ATE^{R} (m)** ↓	**T-ATE^{S} (m)** ↓	**R-ATE (°)** ↓	**T-RTE (m)** ↓	**T-RTE (°)** ↓	**R-RTE (°)** ↓	**AUC@30** ↑	**SR** ↑
DUSt3R [79]	15 / 14	20 / 18	73 / 60	21 / 20	75 / 76	75 / 79	0.12 / 0.07	0.39
MUSt3R [80]	7.8 / 5.7	10 / 3.7	26 / 13	11 / 8.0	37 / 27	31 / 17	0.53 / 0.61	0.81
VGGT [30]	15 / 14	9.9 / 4.5	33 / 15	22 / 20	43 / 32	35 / 21	0.44 / 0.52	0.80
π^3 [81]	15 / 14	4.7 / 1.4	21 / 6.9	21 / 20	26 / 17	24 / 8.5	0.64 / 0.76	0.89
MapAnything [82]	6.1 / 4.2	8.3 / 4.1	30 / 13	8.6 / 6.0	37 / 30	34 / 18	0.50 / 0.59	0.88
DA3 [83]	4.9 / 2.8	6.0 / 2.3	28 / 15	6.9 / 3.9	32 / 23	33 / 20	0.50 / 0.56	0.86
COLMAP [36]	16 / 10	8.1 / 2.3	42 / 10	23 / 14	38 / 25	28 / 12	0.50 / 0.64	0.64
COLMAP^{calib} [36]	10 / 9.7	4.8 / 0.30	15 / 5.0	15 / 13	16 / 7.7	15 / 5.4	0.73 / 0.83	0.87
Best of All	2.8	0.30	5.0	3.9	7.7	5.4	0.83	0.89

described in Appendix Sec. D.

3.6. Data Statistics

At the current stage, the WANDERLAND dataset comprises 530 distinct scenes with over 420,000 frames captured across more than 100 hours of recording, covering a total area of more than 3.8 million square meters. Each scene provides a comprehensive suite of raw and processed data including: synchronized RGB fisheye images with intrinsic calibrations, globally consistent camera poses, colorized metric point clouds, optimized 3D Gaussian Splatting models, extracted collision meshes, and ready-to-simulate USD scenes. To support long-term research growth, we are committed to continuous dataset maintenance and expansion, with an active development roadmap targeting over 1,000 scenes to further enhance diversity and scale for the embodied AI and 3D computer vision community.

4. Experiments

We conduct extensive experiments to answer several critical question at the core of this work:
Q1: Is the best vision-only 3D reconstruction method as good as LIV-SLAM?
Q2: Does geometric grounding lead to a better photorealistic simulation quality?
Q3: Does video-3DGS framework provide reliable environment for training and evaluation?

A1: *Despite recent bursts in vision-only 3D reconstruction, they have large gaps compared to LIV-SLAM and are not reliable in geometric correctness.*

4.1. 3D Reconstruction

Evaluation metrics. For camera pose estimation, we use standard absolute trajectory error (ATE) and relative trajectory error (RTE) to evaluate the translational (T-) and rotational (R-) accuracy of camera poses. Area under curve at 30°(AUC@30) and success rate (SR) evaluates overall camera pose accuracy. See Sec. E.1 for detailed definition.

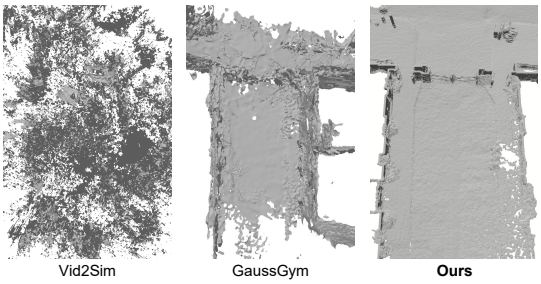

Figure 5. **Mesh qualitative comparison**. All results are reconstructed from the same data in the WANDERLAND dataset. Our framework extracts complete and smooth mesh.

Vision only method produces inaccurate camera pose. In Tab. 2, our evaluation of camera pose estimation against LIV-SLAM ground truth reveals fundamental limitations in vision only methods. In scenes span less than 100 meters, even if we naively take the "best of all" performance, the camera pose estimation accuracy only reaches meter-level metrically. After scale alignment, it still has an average error of 30 cm and 5 degrees. This is due to the inherent modality limitation and the gap is not yet closed by recent foundation models. It prevents the usage of casual videos as data source for reliable reconstruction.

Vision only input prevents accurate mesh extraction. As illustrated in Fig. 5, meshes generated from Vid2Sim [27] and GaussGym [29] exhibit significant noise, fragmentation, and incompletion, stemming from inaccuracies in the underlying neural representations. These deficiencies undermine collision reliability and hinder physical interaction. In contrast, our method directly extracts meshes from the globally consistent LiDAR point cloud, yielding clean, grounded geometry that ensures metric accuracy and completeness for robust simulation. This comparison underscores the necessity of direct geometric sensing for building actionable environments.

A2: *Photorealistic sensor simulation significantly benefits from geometric grounding.*

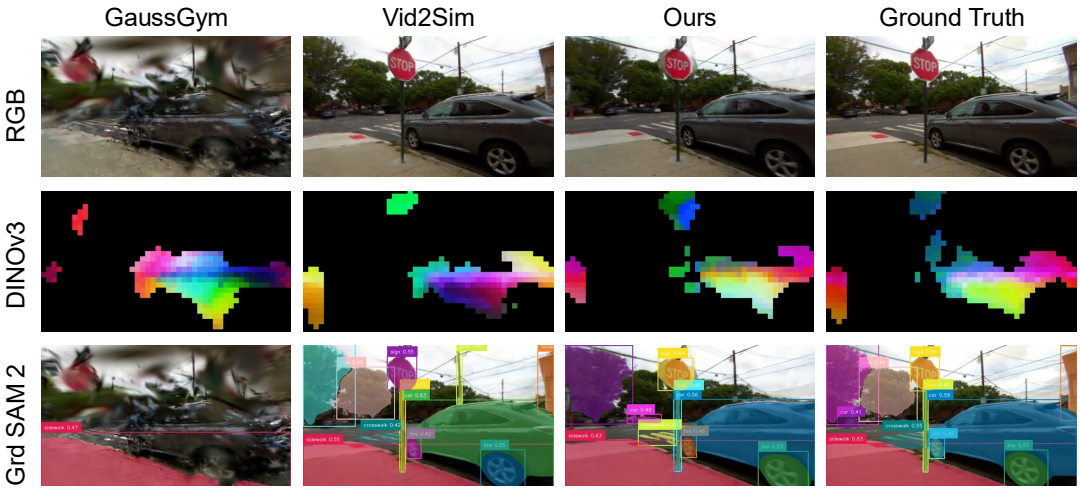

Figure 6. **Photorealism and semantic consistency for sensor simulation**. We show extrapolated view synthesis results from different frameworks, and their inference results from the DINOv3 [85] and Grounded SAM 2 models [86]. The DINOv3 visualization shows the first three PCA components on each patch feature. The foreground is filtered by a small linear classifier on DINOv3 features.

Table 3. **Better NVS from geometric grounding**. All methods are evaluated on our WANDERLAND dataset with the same train and validation split (including both interpolated and extrapolated views.) Vid2Sim [27] uses reconstruction from GLOMAP [28] and GaussGym [29] uses reconstruction from VGGT [30].

Method	Interpolated Views			Extrapolated Views		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
COLMAP Reconstruction						
3DGS [31]	18.27	0.658	0.510	16.90	0.624	0.559
2DGS [88]	17.98	0.593	0.550	16.81	0.631	0.508
3DGUT [89]	18.29	0.654	0.535	17.00	0.619	0.576
Custom Reconstruction						
Vid2Sim [27]	17.20	0.549	0.399	16.49	0.573	0.371
GaussGym [29]	12.17	0.440	0.738	12.63	0.436	0.725
LIV-SLAM Reconstruction						
Ours	20.37	0.688	0.327	17.92	0.591	0.445

4.2. Photorealistic Sensor Simulation

Our method outperforms video based frameworks in novel view synthesis. As shown in Tab. 3, all baseline methods achieve subpar results. GaussGym [29] exhibits the lowest metrics due to the inaccurate 3D reconstruction from VGGT [30]. Vid2Sim's [27] unsatisfactory performance stems from its reliance on inaccurate monocular depth estimation [87] as supervision, introducing additional noise during training. As for other baselines, their limited performance suggests that photometric losses alone fail to leverage geometric scene information effectively. In contrast, our method achieves superior performance by leveraging LIV-SLAM's accurate poses and geometric initialization, demonstrating that robust NVS requires both visual and geometric foundations.

Our framework supports semantically consistent sensor simulation. Beyond standard NVS metrics, we investigate semantic consistency for sensor simulation. As shown in Fig. 6, GaussGym's fragmented renderings prevent reliable segmentation, with Grounded SAM 2 failing to detect critical environmental elements. While Vid2Sim produces more coherent images that support detection and segmentation, its DINOv3 features diverge significantly from ground truth (drastically different PCA colors). This can confuse end-to-end navigation policies that rely on DINO features for semantic understanding [21]. In contrast, our renderings maintain both structural integrity and semantic consistency, enabling accurate segmentation and producing DINOv3 features closely aligned with real imagery. These results demonstrate that rendering quality directly impacts perception models, and that geometric accuracy is essential for reliable embodied AI in open-world environments.

Table 4. **Geometrically grounded environment for RL training**. We compare model performance and their change after post-trained on different simulation environments. Unlike Tab. 5, the results are evaluated on the test split of the WANDERLAND dataset. Numbers inside parentheses indicate metric **change** compared to pretrained model. We use background color to represent model **improvement** and **deterioration** after the RL training.

RL Env.	Methods	NE (m) ↓	SR ↑	SPL ↑	IR ↓
Vid2Sim [27]	NoMaD [90]	19.11 (+4%)	0.24 (0%)	0.23 (-4%)	0.64 (-9%)
	CityWalker [21]	26.40 (+24%)	0.17 (-21%)	0.17 (-19%)	0.64 (-11%)
	MBRA [91]	23.76 (-4%)	0.22 (0%)	0.21 (-5%)	0.69 (+1%)
Ours	NoMaD [90]	17.99 (-2%)	0.26 (+8%)	0.24 (0%)	0.61 (-13%)
	CityWalker [21]	19.02 (-10%)	0.24 (+14%)	0.23 (+14%)	0.55 (-23%)
	MBRA [91]	23.76 (-4%)	0.23 (+5%)	0.23 (+5%)	0.63 (-7%)

A3: *Our framework builds more reliable environment for both training and evaluating embodied navigation.*

4.3. Embodied Navigation

Evaluation metrics. We use standard navigation error (NE), success rate (SR), and success path length (SPL) to evaluate embodied navigation performance. We additionally defined the intervention rate (IR) to evaluate none-

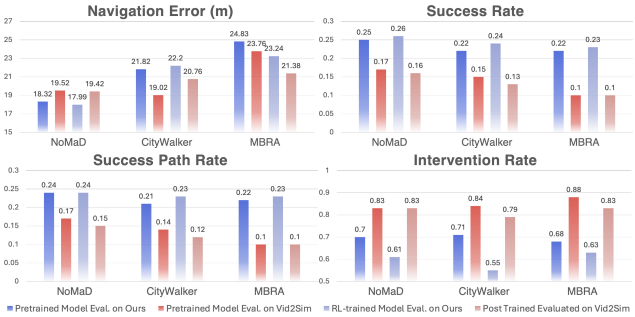

Figure 7. **Geometrically grounded environment for evaluation**. We compare evaluation results of the same models on different environments. Results are evaluated on the WANDERLAND test split. Comparison should be made between different evaluation environments (different colors) for same models.

timeout failure cases. We describe more details on metric definition and experiment setup in Appendix Sec. E.3.

Reinforcement learning with geometric grounding. In Tab. 4, we perform reinforcement learning (RL) training in environments built from different frameworks. Notably, models generally deteriorate after training in environments built by Vid2Sim [27]. Under inaccurate geometry, the RL objective encourages locally shorter but globally unreliable behaviors, which results in lower success rate during evaluation. In contrast, all models significantly improve when trained in our geometrically grounded environments. This comparison highlights that RL fine-tuning is highly sensitive to the underlying simulation fidelity: metrically accurate, collision-consistent geometry is crucial for RL to produce genuinely better navigation policies.

Evaluation with geometric grounding. We further demonstrate that model evaluation can be unreliable in the absence of grounded geometry. As shown in Fig. 7, when evaluated with Vid2Sim-built environments (red bars), all models show a much lower success rate and higher intervention rate compared to our environments (blue bars). This indicates that geometrically unreliable environments fail to support a faithful evaluation. To minimize the sim-to-real gap, geometrically grounded simulation is a must in benchmarking embodied navigation policies.

Benchmarking pretrained navigation policies. Table 5 summarized the performance of pretrained models evaluated on the WANDERLAND dataset. Our first observation is that VLN models generally outperforms point-goal and image-goal models. This is expected as VLN is now considered a less challenging task compared to others (only with the recent help of LLMs). Another observation is that outdoor navigation is generally a more challenging task than indoor scenes. This is due to the longer trajectories, more complex topology, and elevation changes. Overall, the benchmark shows a large research gap in open-world embodied navigation as none of the models reach a success rate over 50%.

Table 5. **Navigation Benchmark**. We benchmark different navigation models for different tasks on our entire WANDERLAND dataset. See Appendix Sec. E.3 for detailed metric definition.

Methods	Indoor Scenes				Outdoor Scenes			
	NE (m) ↓	SR ↑	SPL ↑	IR ↓	NE (m) ↓	SR ↑	SPL ↑	IR ↓
NoMaD [90]	7.04	0.22	0.22	0.52	13.4	0.24	0.24	0.70
CityWalker [21]	7.82	0.39	0.39	0.59	16.4	0.21	0.20	0.72
MBRA [91]	5.28	0.35	0.34	0.55	19.4	0.22	0.22	0.68
NaVid [92]	15.1	0.29	0.25	0.66	28.5	0.15	0.14	0.77
NaVILA [71]	5.13	0.47	0.47	0.41	13.2	0.31	0.31	0.68

5. Discussion and Conclusion

Broader Impact. Beyond bridging the real-to-sim gap for embodied navigation, WANDERLAND provides foundational resources for multiple research domains. The metric-scale camera poses and dense LiDAR point clouds offer unprecedented ground-truth data for foundational vision geometry models. The carefully designed extrapolated views enable rigorous benchmarking of novel view synthesis methods under realistic off-trajectory conditions. Critically, our dataset addresses the scarcity of large-scale metric benchmarks for outdoor 3D vision, enabling new research directions in long-range depth estimation and geometric learning. By providing reliable geometric ground truth at scale, WANDERLAND establishes a new standard for research in 3D computer vision and embodied AI.

Limitations. Our framework has two primary limitations. First, the current capture system operates at 1 FPS due to hardware constraints, resulting in sparser viewpoint sampling than ideal. This limits the density of training views and consequently affects the final rendering quality in highly complex scenes. We plan to address this through hardware upgrades in our ongoing dataset development. Second, while we focus on geometric reconstruction and static environment simulation, real urban environments involve complex dynamics including moving pedestrians, vehicles, and traffic patterns. Modeling these dynamic elements remains an important challenge for future work, requiring integration with behavior prediction and interactive simulation beyond the scope of this paper.

Conclusion. We demonstrate that reliable simulation for open-world embodied AI requires geometrically grounded environments, which is a critical requirement unmet by current video-based 3DGS pipelines. Our work shows that vision-only reconstruction fails to deliver the metric accuracy, reliable geometry, and consistent view synthesis needed for reproducible benchmarking. By introducing the WANDERLAND framework and dataset, we establish a new foundation for embodied AI research, where perception, planning, and evaluation can be built upon accurate and scalable simulations. Moving forward, we argue that geometric grounding is not optional, but essential for the next generation of embodied systems operating in real world.

Acknowledgment

The work was supported in part through NSF grants 2514030, 2238968, and 2345139, in part by NVIDIA Academic Grant Program, and the NYU IT High Performance Computing resources, services, and staff expertise. We thank SkylandX for their technical support. We thank Hellon Luo, Shiqi Wang, Ying Wang, Zhicheng Yang, and Yining Zheng for their help in data collection. We thank Juexiao Zhang and Sihang Li for insightful discussion.

References

[1] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *CVPR*, pages 9339–9347, 2019. 2

[2] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IROS*, pages 7520–7527. IEEE, 2021.

[3] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[4] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

[5] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *CVPR*, pages 11097–11107, 2020.

[6] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 2

[7] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 2, 3

[8] 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. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 3

[10] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *CVPR*, pages 3164–3174, 2020.

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 3

[12] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, pages 9068–9079, 2018. 2, 3

[13] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[15] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[16] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *CVPR*, pages 4122–4134, 2025. 2

[17] Justin Wasserman, Karmesh Yadav, Girish Chowdhary, Abhinav Gupta, and Unnat Jain. Last-mile embodied visual navigation. In *Conference on Robot Learning*, pages 666–678. PMLR, 2023. 2

[18] Joonho Lee, Marko Bjelonic, Alexander Reske, Lorenz Wellhausen, Takahiro Miki, and Marco Hutter. Learning robust autonomous navigation and locomotion for wheeled-legged robots. *Science Robotics*, 9(89):eadi9641, 2024. 2, 3

[19] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, 2016.

[20] Jing Liang, Dibyendu Das, Daeun Song, Md Nahid Hasan Shuvo, Mohammad Durrani, Karthik Taranath, Ivan Penskiy, Dinesh Manocha, and Xuesu Xiao. Gnd: Global navigation dataset with multi-modal perception and multi-category traversability in outdoor campus environments. In *ICRA*, pages 2383–2390. IEEE, 2025. 2

[21] Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang,

and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *CVPR*, pages 6875–6885, 2025. 2, 3, 7, 8

[22] Maks Sorokin, Jie Tan, C Karen Liu, and Sehoon Ha. Learning to navigate sidewalks in outdoor environments. *IEEE Robotics and Automation Letters*, 7(2):3906–3913, 2022. 3

[23] Zhenghao Peng, Zhizheng Liu, and Bolei Zhou. Data-efficient learning from human interventions for mobile robots. *arXiv preprint arXiv:2503.04969*, 2025. 2

[24] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS*, 2021. 3

[25] Jonas Frey, Turcan Tuna, Lanke Frank Tarimo Fu, Cedric Weibel, Katharine Patterson, Benjamin Krummenacher, Matthias Müller, Julian Nubert, Maurice Fallon, Cesar Cadena, and Marco Hutter. Boxi: Design Decisions in the Context of Algorithmic Performance for Robotics. In *Proceedings of Robotics: Science and Systems*, Los Angeles, United States, July 2025. 3

[26] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM TOG*, 36(4):1, 2017. 3

[27] Ziyang Xie, Zhizheng Liu, Zhenghao Peng, Wayne Wu, and Bolei Zhou. Vid2sim: Realistic and interactive simulation from video for urban navigation. In *CVPR*, pages 1581–1591, 2025. 3, 4, 5, 6, 7, 8

[28] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *ECCV*, pages 58–77. Springer, 2024. 2, 3, 7

[29] Alejandro Escontrela, Justin Kerr, Arthur Allshire, Jonas Frey, Rocky Duan, Carmelo Sferrazza, and Pieter Abbeel. Gaussgym: An open-source real-to-sim framework for learning locomotion from pixels. *arXiv preprint arXiv:2510.15352*, 2025. 3, 4, 6, 7

[30] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 2, 3, 6, 7

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023. 2, 7

[32] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022. 2, 3

[33] Xiaohan Lei, Min Wang, Wengang Zhou, and Houqiang Li. Gaussnav: Gaussian splatting for visual navigation. *IEEE TPAMI*, 2025. 3

[34] Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. *IEEE Transactions on Robotics*, 2025. 3

[35] Gunjan Chhablani, Xiaomeng Ye, Muhammad Zubair Irshad, and Zsolt Kira. Embodiedsplat: Personalized real-to-sim-to-real navigation with gaussian splats from a mobile device. *arXiv preprint arXiv:2509.17430*, 2025. 2

[36] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 2, 3, 6

[37] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, pages 71–91. Springer, 2024. 2

[38] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, pages 5354–5363, 2024. 2, 3, 5

[39] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM TOG*, 43(6):1–13, 2024. 2, 3

[40] Xiangyu Han, Zhen Jia, Boyi Li, Yan Wang, Boris Ivanovic, Yurong You, Lingjie Liu, Yue Wang, Marco Pavone, Chen Feng, and Yiming Li. Extrapolated urban view synthesis benchmark. In *ICCV*, 2025. 2

[41] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *CVPR*, pages 4369–4379, 2023. 3

[42] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024. 3

[43] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, pages 197–214. Springer, 2024. 3

[44] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 3

[45] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *ICRA*, pages 15–22. IEEE, 2014.

[46] Shinya Sumikura, Mikiya Shibuya, and Ken Sakurada. Openvslam: A versatile visual slam framework. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2292–2295, 2019.

[47] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *NeurIPS*, 36:39033–39051, 2023.

[48] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *ECCV*, pages 424–440. Springer, 2024.

[49] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 34(4):1004–1020, 2018. 3, 4

[50] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IROS*, pages 4758–4765. IEEE, 2018. 3

[51] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *IROS*, pages 5135–5142. IEEE, 2020.

[52] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022.

[53] Chao Chen, Xinhao Liu, Yiming Li, Li Ding, and Chen Feng. Deepmapping2: Self-supervised large-scale lidar map optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9316, 2023.

[54] Tiziano Guadagnino, Benedikt Mersch, Saurabh Gupta, Ignacio Vizzo, Giorgio Grisetti, and Cyrill Stachniss. Kiss-slam: A simple, robust, and accurate 3d lidar slam system with enhanced generalization capabilities. *arXiv preprint arXiv:2503.12660*, 2025. 3

[55] Chunran Zheng, Wei Xu, Zuhao Zou, Tong Hua, Chongjian Yuan, Dongjiao He, Bingyang Zhou, Zheng Liu, Jiarong Lin, Fangcheng Zhu, et al. Fast-livo2: Fast, direct lidar-inertial-visual odometry. *IEEE Transactions on Robotics*, 2024. 3, 4

[56] Jiarong Lin and Fu Zhang. R3live: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package. In *ICRA*, pages 10672–10678. IEEE, 2022.

[57] Bonan Liu, Guoyang Zhao, Jianhao Jiao, Guang Cai, Chengyang Li, Handi Yin, Yuyang Wang, Ming Liu, and Pan Hui. Omnicolor: a global camera pose optimization approach of lidar-360camera fusion for colorizing point clouds. In *ICRA*, pages 6396–6402. IEEE, 2024. 3

[58] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, pages 12–22, 2023. 3

[59] Arunkumar Byravan, Jan Humplik, Leonard Hasenclever, Arthur Brussee, Francesco Nori, Tuomas Haarnoja, Ben Moran, Steven Bohez, Fereshteh Sadeghi, Bojan Vujatovic, et al. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields. In *ICRA*, pages 9362–9369, 2023. 3

[60] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 3

[61] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, 2020.

[62] Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, and Devendra Singh Chaplot. Navigating to objects specified by images. In *CVPR*, pages 10916–10925, 2023. 3

[63] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 3

[64] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. In *ICLR*, 2018. 3

[65] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *CVPR*, pages 12875–12884, 2020.

[66] Juexiao Zhang, Gao Zhu, Sihang Li, Xinhao Liu, Haorui Song, Xinran Tang, and Chen Feng. Multiview scene graph. *NeurIPS*, 37:17761–17788, 2024. 3

[67] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. 3, 5

[68] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, pages 4392–4412, 2020. 5

[69] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, pages 104–120. Springer, 2020. 3

[70] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *ICCV*, 2021. 4

[71] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. In *RSS*, 2025. 4, 8

[72] Shaozu Cao, Xiuyuan Lu, and Shaojie Shen. Gvins: Tightly coupled gnss–visual–inertial fusion for smooth and consistent state estimation. *IEEE Transactions on Robotics*, 38(4):2004–2021, 2022. 4

[73] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, et al. Egoblur: Responsible innovation in aria. *arXiv preprint arXiv:2308.13093*, 2023. 4

[74] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 4

[75] Zimu Liao, Siyan Chen, Rong Fu, Yi Wang, Zhongling Su, Hao Luo, Li Ma, Linning Xu, Bo Dai, Hengjie Li, et al. Fisheye-gs: Lightweight and extensible gaussian splatting module for fisheye cameras. *arXiv preprint arXiv:2409.04751*, 2024. 4

[76] Youming Deng, Wenqi Xian, Guandao Yang, Leonidas Guibas, Gordon Wetzstein, Steve Marschner, and Paul Debevec. Self-calibrating gaussian splatting for large field of view reconstruction. In *ICCV*, 2025. 4

[77] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *CVPR*, pages 26024–26035, 2025. 5, 13

[78] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, page 163–169. Association for Computing Machinery, 1987. 5, 15

[79] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024. 6

[80] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *CVPR*, pages 1050–1060, 2025. 6

[81] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv e-prints*, pages arXiv–2507, 2025. 6

[82] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 6

[83] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 6

[84] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5, 15

[85] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 7

[86] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 7

[87] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 37:21875–21911, 2024. 7

[88] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 7

[89] Qi Wu, Janick Martinez Esturo, Ashkan Mirzaei, Nicolas Moenne-Loccoz, and Zan Gojcic. 3dgut: Enabling distorted cameras and secondary rays in gaussian splatting. In *CVPR*, pages 26036–26046, 2025. 7

[90] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *ICRA*, pages 63–70. IEEE, 2024. 7, 8

[91] Noriaki Hirose, Lydia Ignatova, Kyle Stachowicz, Catherine Glossop, Sergey Levine, and Dhruv Shah. Learning to drive anywhere with model-based reannotation. *arXiv preprint arXiv:2505.05592*, 2025. 7, 8

[92] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024. 8

[93] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 13

[94] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 18