OmniWorld: A MULTI-DOMAIN AND MULTI-MODAL DATASET FOR 4D WORLD MODELING

Yang Zhou[1,2] **Yifan Wang**[2,3] **Jianjun Zhou**[2,4,5] **Wenzheng Chang**[2,3] **Haoyu Guo**[2] **Zizun Li**[2,6] **Kaijing Ma**[1,2] **Xinyue Li**[2,3] **Yating Wang**[2,7] **Haoyi Zhu**[2,6] **Mingyu Liu**[2,4] **Dingning Liu**[1,2] **Jiange Yang**[2,8] **Zhoujie Fu**[2,9] **Junyi Chen**[2,3] **Chunhua Shen**[4] **Jiangmiao Pang**[2] **Kaipeng Zhang**[2] **Tong He**[2,5,†]

[1]Fudan University [2]Shanghai AI Laboratory [3]Shanghai Jiao Tong University [4]Zhejiang University [5]Shanghai Innovation Institute [6]University of Science and Technology of China [7]Tongji University [8]Nanjing University [9]Nanyang Technological University [†]Corresponding Author

Figure 1: We introduce ***OmniWorld***, a large-scale, multi-domain, and multi-modal dataset. *OmniWorld* provides a rich resource for 4D world modeling by integrating high-quality data from multiple domains and offers a variety of data types, including depth maps, camera poses, text captions, optical flow and foreground masks. *OmniWorld* is designed to accelerate the development of more general models for modeling the real physical world.

ABSTRACT

The field of 4D world modeling—aiming to jointly capture spatial geometry and temporal dynamics—has witnessed remarkable progress in recent years, driven by advances in large-scale generative models and multimodal learning. However, the development of truly general 4D world models remains fundamentally constrained by the availability of high-quality data. Existing datasets and benchmarks often lack the dynamic complexity, multi-domain diversity, and spatial-temporal annotations required to support key tasks such as 4D geometric reconstruction, future prediction, and camera-controlled video generation. To address this gap, we introduce *OmniWorld*, a large-scale, multi-domain, multi-modal dataset specifically designed for 4D world modeling. *OmniWorld* consists of a newly collected *OmniWorld-Game* dataset and several curated public datasets spanning diverse domains. Compared with existing synthetic datasets, *OmniWorld-Game* provides richer modality coverage, larger scale, and more realistic dynamic interactions. Based on this dataset, we establish a challenging benchmark that exposes the limitations of current state-of-the-art (SOTA) approaches in modeling complex 4D environments. Moreover, fine-tuning existing SOTA methods on *OmniWorld* leads to significant performance gains across 4D reconstruction and video generation tasks, strongly validating *OmniWorld* as a powerful resource for training and evaluation. We envision *OmniWorld* as a catalyst for accelerating the development of general-purpose 4D world models, ultimately advancing machines' holistic understanding of the physical world. Project Page: `https://yangzhou24.github.io/OmniWorld/`

Table 1: **Comparisons between *OmniWorld-Game* and existing synthetic datasets.** *OmniWorld-Game* surpasses existing public synthetic datasets in modal diversity and data scale.

Dataset	Scene Type	Motion	Resolution	# Frames	Data modality				
					Depth	Camera	Text	Optical flow	Fg. masks
MPI Sintel (Butler et al., 2012)	Mixed	Dynamic	1024 × 436	1K	✔	✔	✗	✔	✔
FlyingThings++ (Mayer et al., 2016; Harley et al., 2022)	Outdoor	Dynamic	960 × 540	28K	✔	✗	✗	✔	✔
TartanAir (Wang et al., 2020)	Mixed	Dynamic	640 × 480	1,000K	✔	✔	✗	✔	✔
BlendedMVS (Yao et al., 2020)	Mixed	Static	768 × 576	17K	✔	✔	✗	✗	✗
HyperSim (Roberts et al., 2021)	Indoor	Static	1024 × 768	77K	✔	✔	✗	✗	✔
Dynamic Replica (Karaev et al., 2023)	Indoor	Dynamic	1280 × 720	169K	✔	✔	✗	✔	✔
Spring (Mehl et al., 2023)	Mixed	Dynamic	1920 × 1080	23K	✔	✔	✗	✔	✗
EDEN (Le et al., 2021)	Outdoor	Static	640 × 480	300K	✔	✔	✗	✔	✔
PointOdyssey (Zheng et al., 2023)	Mixed	Dynamic	960 × 540	216K	✔	✔	✗	✗	✔
SeKai-Game (Li et al., 2025)	Outdoor	Dynamic	1920 × 1080	4,320K	✗	✔	✔	✗	✗
***OmniWorld-Game* (Ours)**	Mixed	Dynamic	1280 × 720	**18,515K**	✔	✔	✔	✔	✔

1 INTRODUCTION

The development of world models (DeepMind, 2025; Ha & Schmidhuber, 2018; Agarwal et al., 2025; LeCun, 2022; Hafner et al., 2023) has become a central pursuit in visual intelligence systems, aiming to build systems that can simulate and reason about the physical world. This capability goes beyond simple static perception, demanding models that can simulate dynamic environments, predict object motion, infer causality, and generate content that adheres to physical laws. Such spatio-temporal modeling is a cornerstone for effective world models, with its development critically dependent on large-scale, multi-domain, and multi-modal datasets (Feng et al., 2024; Team et al., 2025a; Chen et al., 2025; He et al., 2025b; Team et al., 2025b; Yu et al., 2025b;a).

Two fundamental tasks that reflect a model's world modeling capability have drawn widespread attention: 3D geometric foundation models (Wang et al., 2024c; Leroy et al., 2024; Zhang et al., 2024; Yang et al., 2025a; Tang et al., 2024; Wang et al., 2025b; Zhang et al., 2025; Wang et al., 2025a;d), and camera-controlled video generation models (Wang et al., 2024d; He et al., 2024; Zheng et al., 2024; Bahmani et al., 2024; Bai et al., 2025; YU et al., 2025). The former aims to extract comprehensive 3D geometric information from 2D image inputs, while the latter focuses on generating dynamic video content that follows precise spatio-temporal instructions. Both tasks heavily rely on large-scale, high-quality datasets with rich modalities, including RGB images, depth maps, and camera poses.

However, existing benchmarks and datasets for evaluating and training these models have significant limitations. In the domain of 3D geometric foundation models, existing benchmarks suffer from short sequence lengths, which constrain the evaluation of a model's long-term robustness. For example, Sintel (Butler et al., 2012), which is a widely used dataset, consists of videos with an average length of only 50 frames. Furthermore, the limited motion amplitude and single-action types within these datasets (e.g., Bonn's (Palazzolo et al., 2019) focuses on indoor human motion, Kitti's (Geiger et al., 2013) focuses on outdoor street scenes) fail to comprehensively evaluate model performance in complex, dynamic environments. Similarly, in the field of camera-controlled video generation, mainstream datasets like RealEstate10K (Zhou et al., 2018) primarily consist of static scenes with smooth camera trajectories. This lack of diverse object motion and complex camera operations results in a noticeable gap between the dataset's content and real-world scenarios, thereby hindering a comprehensive assessment of a model's true capabilities.

From the perspective of training data, there is a critical scarcity of high-quality, multi-domain, multi-modal datasets that include rich geometric annotations. For instance, in image or video generation, while there are numerous image-text (Schuhmann et al., 2022; Gadre et al., 2023) or video-text datasets (Chen et al., 2024; Nan et al., 2024; Ju et al., 2024), they often lack critical geometric modalities such as depth maps, camera poses, and optical flow. Similarly, the demand for large-scale, diverse datasets with accurate geometric annotations is increasingly urgent for 3D geometric foundation models.

To address these shortcomings, we introduce *OmniWorld*, a large-scale, multi-domain, and multi-modal dataset composed of a self-collected high-quality *OmniWorld-Game* synthetic dataset and several public datasets. Its core characteristics are: **1) High-Quality 4D Data.** *OmniWorld-Game* is a massive synthetic video dataset comprising over 96K clips and more than 18M frames, with a total duration of over 214 hours. It is captured from diverse game environments with 720P RGB images, dense ground truth depth maps, accurate camera poses, and annotations for text captions, optical flow and foreground masks. As shown in Tab. 1, the dataset significantly surpasses existing

Table 2: ***OmniWorld* structure**. A smiling face (🙂) indicates the modality is newly (re-)annotated by us, a green check (✔) denotes ground-truth data that already exists in the original dataset, and a red cross (✗) marks missing modalities.

Dataset	Domain	# Seq.	FPS	Resolution	# Frames	Data modality				
						Depth	Camera	Text	Opt. flow	Fg. masks
OmniWorld-Game	Simulator	96K	24	1280×720	18,515K	🙂	🙂	🙂	🙂	🙂
AgiBot (Bu et al., 2025)	Robot	20K	30	640×480	39,247K	🙂	✔	✔	✗	🙂
DROID (Khazatsky et al., 2024)	Robot	35K	60	1280×720	26,643K	🙂	✔	🙂	🙂	🙂
RH20T (Fang et al., 2024)	Robot	109K	10	640×360	53,453K	✗	✔	🙂	🙂	🙂
RH20T-Human (Fang et al., 2024)	Human	73K	10	640×360	8,875K	✗	✔	🙂	✗	✗
HOI4D (Liu et al., 2022)	Human	2K	15	1920×1080	891K	🙂	🙂	🙂	🙂	✔
Epic-Kitchens (Damen et al., 2018)	Human	15K	30	1280×720	3,635K	✗	🙂	🙂	✗	✗
Ego-Exo4D (Grauman et al., 2024)	Human	4K	30	1024×1024	9,190K	✗	✔	🙂	🙂	✗
HoloAssist (Wang et al., 2023)	Human	1K	30	896×504	13,037K	✗	🙂	🙂	🙂	✗
Assembly101 (Sener et al., 2022)	Human	4K	60	1920×1080	110,831K	✗	✔	🙂	🙂	🙂
EgoDex (Hoque et al., 2025)	Human	242K	30	1920×1080	76,631K	✗	✔	🙂	✗	✗
CityWalk (Li et al., 2025)	Internet	7K	30	1280×720	13,096K	✗	🙂	✔	✗	✗

public synthetic datasets in modal diversity and scale. **2) Multi-Domain Coverage.** By integrating datasets from four key domains including simulator, robot, human, and the internet, *OmniWorld* covers a wide range of real-world and virtual scenarios, greatly enhancing data diversity. **3) Multi-Modality Annotations.** *OmniWorld* provides a rich suite of multi-modal annotations, crucial for detailed world modeling, as shown in Tab. 2.

Based on *OmniWorld-Game*, we propose a new benchmark for both 3D geometric foundation models and camera-controlled video generation models. Our *OmniWorld-Game* benchmark provides challenging, complex scenarios and dynamics that accurately reflect a model's true world capabilities, revealing the limitations of current SOTAs. By fine-tuning existing SOTAs (e.g., DUSt3R (Wang et al., 2024c), CUT3R (Wang et al., 2025b), Reloc3r (Dong et al., 2024), AC3D (Bahmani et al., 2024)) with *OmniWorld*, we demonstrate significant performance improvements on public benchmarks. This strongly validates *OmniWorld* as a powerful training resource for enhancing world modeling capabilities.

In summary, our contributions are as follows:

1. We introduce *OmniWorld*, a multi-domain and multi-modal dataset designed to address the lack of diversity in existing datasets. Its self-collected subset, *OmniWorld-Game*, surpasses current synthetic datasets in both modality diversity and data volume.
2. We establish a comprehensive benchmark for 3D geometric foundation models and camera-controlled video generation models based on *OmniWorld-Game*, providing a unified platform for evaluation.
3. We fine-tune several SOTAs on *OmniWorld* and observe significant performance gains, underscoring its value as a training resource.

2 *OmniWorld* DATASET

2.1 DATA ACQUISITION

Our data acquisition strategy is centered on our novel, self-collected *OmniWorld-Game* dataset, which is strategically supplemented with curated data from three other distinct domains: robot, human, and internet, as illustrated in Fig. 2. This strategy allows us to integrate the strengths of diverse data sources to comprehensively capture real-world complexity.

Simulator domain. To acquire the high-precision and temporally consistent multi-modal data that is hard to obtain in the real world, we collect *OmniWorld-Game* from game environments. Following prior works (Richter et al., 2016; Yang et al., 2024a; Feng et al., 2024; Team et al., 2025a), we utilize ReShade (ReShade Contributors, 2024) to access depth information during the rendering process, and simultaneously capture synchronized RGB images from the screen using OBS (Contributors, 2024). This approach offers significant advantages: 1) High-Precision Modal Data. We can precisely control the environment and acquire accurate depth data, which is often unattainable

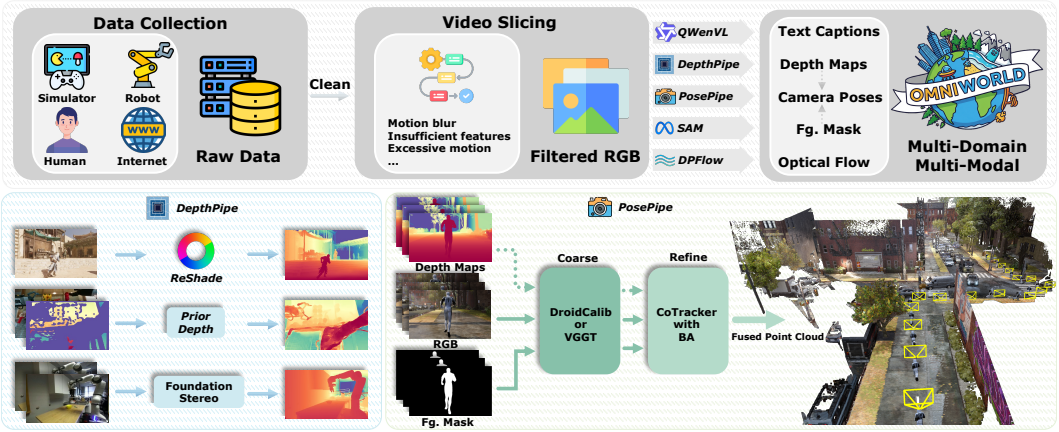


Figure 2: ***OmniWorld* acquisition and annotation pipeline.** We collect raw data from diverse domains and apply a video slicing filter to obtain high-quality RGB sequences. These sequences are then processed through a suite of specialized pipelines to generate multi-modal annotations, including text captions, depth maps, camera poses, foreground masks, and optical flow.

in real-world settings. 2) Rich Real-World Scene Simulation. Modern virtual environments provide highly realistic graphics and diverse simulations of real-world scenarios, such as complex settings from wilderness to urban areas.

Robot domain. We integrate public datasets from robot manipulation and human-robot interaction tasks, including AgiBot (Bu et al., 2025), DROID (Khazatsky et al., 2024), and RH20T (Fang et al., 2024). These datasets provide valuable sequences of robot-environment interactions and navigation, which are essential for tasks involving robotic manipulation and physical world understanding.

Human domain. We incorporate public datasets describing various human activities, including RH20T-Human (Fang et al., 2024), HOI4D (Liu et al., 2022), Epic-Kitchens (Damen et al., 2018), Ego-Exo4D (Grauman et al., 2024), HoloAssist (Wang et al., 2023), Assembly101 (Sener et al., 2022), and EgoDex (Hoque et al., 2025). These datasets capture diverse human behaviors, ranging from daily activities to complex assembly tasks, from both egocentric and exocentric perspectives.

Internet domain. To acquire large-scale, realistic, and diverse in-the-wild scene data, we utilize the CityWalk dataset (Li et al., 2025). It offers rich real-world street view videos from the internet.

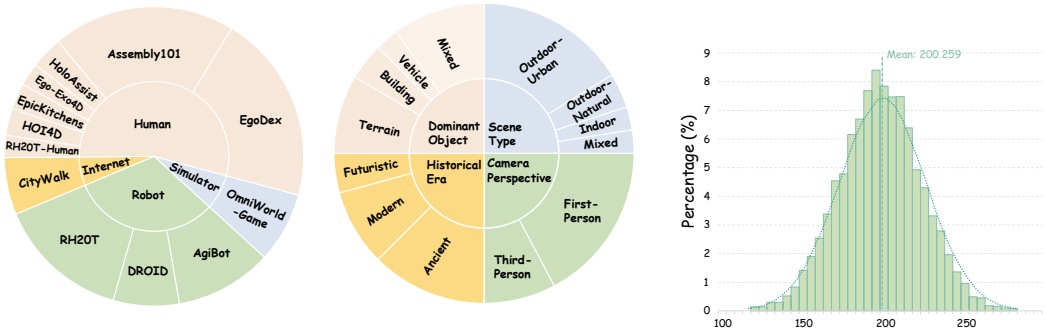
To prepare the raw data, we perform video slicing to ensure high quality and temporal coherence. This crucial preprocessing step filters out unsuitable frames (e.g., those with motion blur or insufficient features) and segments long recordings into shorter, manageable clips. The resulting high-quality video segments are then passed to our multi-modal annotation pipeline.

2.2 Data Annotation

We primarily annotate the following key modalities: depth maps, camera poses, text captions, optical flow, and foreground masks (see Fig. 2 for the overall pipeline). Here we briefly introduce the annotation method of each modality.

Depth maps. Accurate depth information is paramount for geometric modeling. To ensure the quality and consistency of depth maps, we adopt a tailored approach based on the data source. For the self-collected dataset *OmniWorld-Game*, as mentioned in Sec. 2.1, we directly access depth information during the rendering process using tools like ReShade (ReShade Contributors, 2024).

For public datasets like AgiBot (Bu et al., 2025) and HOI4D (Liu et al., 2022), which often provide noisy and sparse raw depth maps, we employ Prior Depth Anything (Wang et al., 2025e) to robustly optimize them, yielding denser and more reliable depth annotations. For the public stereo dataset DROID (Khazatsky et al., 2024), we leverage FoundationStereo (Wen et al., 2025) for stereo depth estimation on this dataset.



(a) *OmniWorld* Compositional Distribution

(b) *OmniWorld-Game* Internal Composition

(c) Caption Tokens Distribution

Figure 3: **Statistical information of *OmniWorld*.** (a) displays compositional distribution of data from different domains within *OmniWorld*, (b) presents internal composition of *OmniWorld-Game*. (c) shows caption tokens distribution of *OmniWorld*.

Foreground masks. To provide precise, temporally consistent masks of primary subjects, we develop specialized automated pipelines. For robot domain data, we use RoboEngine (Yuan et al., 2025) to generate initial masks for keyframes, followed by temporal tracking and fusion with SAM 2 (Ravi et al., 2024). For *OmniWorld-Game* (e.g., player characters in third-person view), we leverage Grounding DINO (Liu et al., 2023) to detect initial bounding boxes within predefined regions of keyframes, which then serve as prompts for SAM (Kirillov et al., 2023). These generated masks can be used as dynamic foreground masks to guide camera pose estimation.

Camera poses. Accurate camera pose annotation in dynamic videos is highly challenging due to transitions, weakly textured areas, and abrupt movements that hinder traditional Structure-from-Motion methods (Rockwell et al., 2025; Li et al., 2024). Following prior work (Team et al., 2025a), we develop a robust, automated, two-stage pipeline for dynamic camera pose annotation, whose principles are validated across diverse data types.

The pipeline leverages the pre-computed foreground masks to focus on static background regions. The stages include: 1) Coarse camera pose estimation leveraging VGGT (Wang et al., 2025a) for videos without depth or DroidCalib (Hagemann et al., 2023) with depth constraints; 2) Camera pose refinement through dense point tracking (SIFT (Lowe, 2004), SuperPoint (DeTone et al., 2018) with CoTracker3 (Karaev et al., 2024)) on static regions and subsequent bundle adjustment to minimize reprojection errors, optionally enhanced by forward-backward reprojection with depth information (Chen et al., 2019).

Text captions. We generate text descriptions using a semi-automated approach centered on the Qwen2-VL-72B-Instruct model (Wang et al., 2024a). We employ domain-specific prompting strategies for each 81-frame video segment. For instance, in the *OmniWorld-Game* domain, we generate multi-faceted descriptions covering different viewpoints (e.g., first- and third-person), character actions, background details, and camera movements.

Optical flow. We generate optical flow annotations using DPFlow (Morimitsu et al., 2025) to capture dense, pixel-level motion. Unlike models that require downsampling high-resolution inputs (Teed & Deng, 2020), DPFlow processes videos at their original resolution. This makes it ideal for our high resolutions dataset.

2.3 Data Statistics

OmniWorld is a large-scale dataset composed of 12 distinct datasets from four domains: simulators, robots, humans, and the internet (see Tab. 2 for a summary). It contains over 600K video sequences and 300M frames with high resolutions. The dataset is richly annotated with multiple modalities, including depth, camera poses, text, optical flow, and foreground masks.

As shown in Fig. 3a, the human domain constitutes the largest portion of *OmniWorld*, highlighting its focus on real-world activities. Our self-collected *OmniWorld-Game* subset is particularly diverse, as detailed in Fig. 3b. It spans various scene types (e.g., outdoor-urban, indoor), camera perspectives

Table 3: **Monocular depth & video depth estimation** on *OmniWorld-Game* benchmark.

Method	Mono-Depth		Video-Depth				FPS
	scale		scale		scale&shift		
	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	
DUSt3R (Wang et al., 2024c)	0.742	0.460	0.709	0.447	0.379	0.560	0.96
MASt3R (Leroy et al., 2024)	0.485	0.560	<u>0.482</u>	<u>0.579</u>	<u>0.217</u>	<u>0.724</u>	0.79
MonST3R (Zhang et al., 2024)	0.670	0.493	0.669	0.505	0.272	0.648	0.95
Fast3R (Yang et al., 2025a)	0.755	0.404	0.741	0.384	0.464	0.531	<u>14.99</u>
CUT3R (Wang et al., 2025b)	0.624	0.518	0.690	0.479	0.429	0.603	10.75
FLARE (Zhang et al., 2025)	0.664	0.475	0.757	0.453	0.511	0.527	4.24
VGGT (Wang et al., 2025a)	0.531	0.554	**0.440**	**0.625**	**0.194**	**0.755**	**18.75**
MoGe-1 (Wang et al., 2024b)	<u>0.459</u>	<u>0.586</u>	–	–	–	–	–
MoGe-2 (Wang et al., 2025c)	**0.401**	**0.589**	–	–	–	–	–

(first-person and third-person), historical eras (ancient to futuristic), and dominant objects (natural terrain, architecture, vehicles). This multi-dimensional diversity ensures the data is both challenging and comprehensive.

Furthermore, *OmniWorld* features structured and detailed text annotations. The captions typically range from 150 to 250 tokens (Fig. 3c), a density that significantly surpasses other large-scale video-text datasets like OpenVid-1M (Nan et al., 2024) and Panda-70M (Chen et al., 2024).

3 *OmniWorld-Game* BENCHMARK

3.1 3D GEOMETRIC PREDICTION BENCHMARK

Benchmark design. Existing benchmarks for 3D Geometric Foundation Models (GFMs) often feature short sequences and limited motion dynamics. For example, Sintel (Butler et al., 2012) sequences average only 50 frames, while datasets like Bonn (Palazzolo et al., 2019) and KITTI (Geiger et al., 2013) are confined to specific scenarios (e.g., indoor human motion or outdoor street views). These limitations hinder the comprehensive evaluation of a model's long-term and complex-scene modeling capabilities. To address this, our *OmniWorld-Game* benchmark provides a more challenging testbed featuring extended, high-resolution sequences (up to 384 frames at 720P) with diverse and complex motions.

Evaluation details. We evaluate a suite of recent GFMs, including DUSt3R (Wang et al., 2024c), MASt3R (Leroy et al., 2024), MonST3R (Zhang et al., 2024), Fast3R (Yang et al., 2025a), CUT3R (Wang et al., 2025b), FLARE (Zhang et al., 2025), VGGT (Wang et al., 2025a), and MoGe (Wang et al., 2024b; 2025c). The evaluation is conducted on two tasks: monocular depth estimation and video depth estimation.

Analysis. Our evaluation on *OmniWorld-Game* reveals significant challenges for current GFMs (Tab 3). For monocular depth estimation, MoGe-2 (Wang et al., 2025c) achieves the best quantitative results. This finding is supported by qualitative results in the supplementary materials, where it produces visibly sharper depth maps. In the more demanding video depth estimation task, VGGT (Wang et al., 2025a) demonstrates superior accuracy and efficiency (FPS). Point cloud visualizations (Fig. 4) confirm that VGGT generates more coherent 3D structures, yet even it produces noticeable artifacts in highly dynamic scenes. No single GFM masters all tasks on *OmniWorld-Game*. The results highlight that current SOTA models still struggle with long-sequence consistency and complex dynamics, validating our benchmark as a challenging testbed for advancing future research.

3.2 CAMERA-CONTROLLED VIDEO GENERATION BENCHMARK

Benchmark design. Existing benchmarks for camera-controlled video generation, such as RealEstate10K (Zhou et al., 2018), are often limited to static scenes with smooth camera paths, failing to reflect real-world complexity. In contrast, our *OmniWorld-Game* benchmark provides a more challenging evaluation environment, featuring rich dynamic content, complex camera trajectories, and diverse scenes.

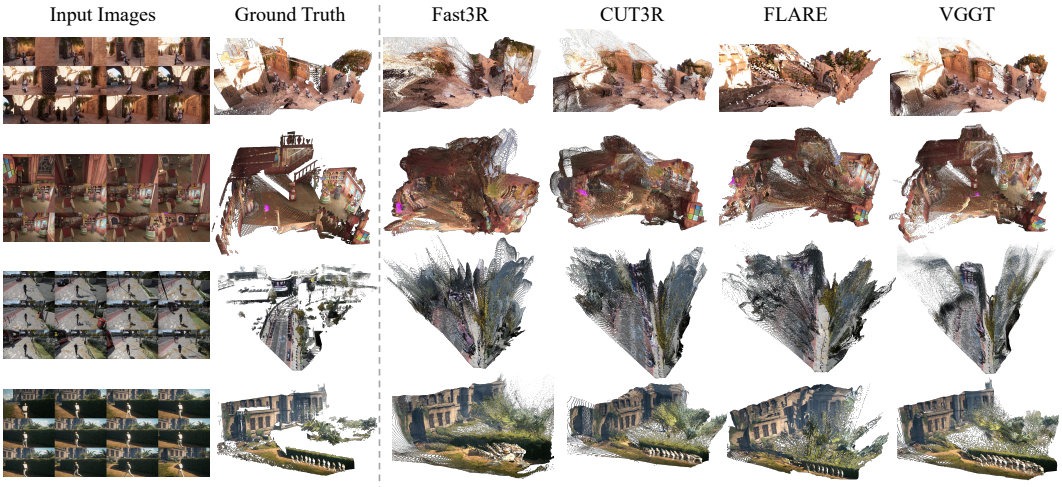


Figure 4: **Qualitative comparison of multi-view 3D reconstruction** on *OmniWorld-Game* benchmark.

Table 4: **Camera-controlled video generation evaluation** on *OmniWorld-Game* benchmark.

Method	TransErr↓	RotErr↓	CamMC↓	FVD	
				VideoGPT↓	StyleGAN↓
AC3D (T2V) (Bahmani et al., 2024)	6.2788	0.8867	6.6965	1745.778	1594.885
MotionCtrl (I2V) (Wang et al., 2024d)	7.8633	1.1402	8.2710	694.342	745.652
CamCtrl (I2V) (He et al., 2024)	**1.2882**	**0.2022**	**1.3856**	**615.417**	**637.574**
CAMI2V (I2V) (Zheng et al., 2024)	5.9626	0.5087	6.2010	837.185	742.594

Evaluation details. We evaluate several recent models, including the Text-to-Video model AC3D (Bahmani et al., 2024) and Image-to-Video models like CamCtrl (He et al., 2024), MotionCtrl (Wang et al., 2024d), CAMI2V (Zheng et al., 2024). Following prior work (Zheng et al., 2024), we assess performance using two sets of metrics: camera-controlled accuracy (RotError, TransError, CamMC) and perceptual quality (Fréchet Video Distance, FVD) (Unterthiner et al., 2018).

Analysis. Our analysis on *OmniWorld-Game* reveals current models failing to achieve either high generation quality or precise camera control. For instance, the Text-to-Video model AC3D (Bahmani et al., 2024) generates subtle dynamics and fails to follow camera paths, resulting in poor quantitative and qualitative scores (Tab. 4, Fig. 5). Among Image-to-Video (I2V) models, CamCtrl (He et al., 2024) shows better quantitative performance. However, its generated videos often suffer from blurry moving characters as shown in Fig. 5. Other methods, including MotionCtrl (Wang et al., 2024d) and CAMI2V (Zheng et al., 2024), face similar quality degradation issues. These results underscore the unique challenges posed by our benchmark in evaluating spatio-temporal generation capabilities.

4 Model Fine-tuning and Efficacy Validation

4.1 Improving 3D Geometric Prediction with *OmniWorld*

To demonstrate *OmniWorld*'s value as a training resource, we fine-tune three baseline models, DUSt3R (Wang et al., 2024c), CUT3R (Wang et al., 2025b), and Reloc3r (Dong et al., 2024), on subsets of our dataset. The fine-tuned models consistently surpass their original performance across monocular depth estimation (Tab. 5), video depth estimation (Tab. 6), and camera pose estimation (see supplementary). Notably, for monocular depth, the fine-tuned DUSt3R not only improved upon its baseline but also outperformed MonST3R Zhang et al. (2024), which is fine-tuned on several existing dynamic datasets. The enhancements in video depth estimation also underscore *OmniWorld*'s effectiveness in improving temporal consistency. These results validate that *OmniWorld*'s scale and diversity provide a powerful resource for boosting the generalization and robustness of 3D geometric foundation models.

Figure 5: **Qualitative comparison of camera-controlled video generation** on *OmniWorld-Game* benchmark. AC3D takes text as a condition signal. MotionCtrl, CamCtrl, CAMI2V take an image as a condition signal. Condition images are the first images of each row.

Table 5: **Comparison of original and fine-tuned models for monocular depth estimation** on Sintel (Butler et al., 2012), Bonn (Palazzolo et al., 2019), KITTI (Geiger et al., 2013) and NYU-v2 (Silberman et al., 2012). * denotes models that have been fine-tuned on *OmniWorld*.

Method	Sintel		Bonn		KITTI		NYU-v2	
	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
DUSt3R (Wang et al., 2024c)	0.488	**0.532**	0.139	0.831	0.109	0.873	**0.081**	**0.909**
MonST3R (Zhang et al., 2024)	0.402	0.525	0.069	**0.954**	0.098	0.895	0.094	0.887
DUSt3R*	**0.370**	0.529	**0.067**	0.948	**0.088**	**0.932**	0.089	0.902
CUT3R (Wang et al., 2025b)	0.420	0.520	**0.058**	**0.967**	0.097	0.914	0.081	0.914
CUT3R*	**0.408**	**0.522**	0.075	0.944	**0.087**	**0.935**	**0.075**	**0.920**

4.2 Enhancing Camera-Controlled Video Generation with *OmniWorld*

Existing datasets for camera-controlled video generation, such as RealEstate10K (Zhou et al., 2018), are often limited to static scenes with simple camera movements, which restricts a model's ability to handle dynamic content. To address this data bottleneck, we fine-tuned the AC3D (Bahmani et al., 2024) baseline on *OmniWorld*. This approach aligns with prior findings (He et al., 2025a) that highlight the critical role of dynamic data in improving camera control. As shown in Tab. 7, our fine-tuned model significantly outperforms the original baseline on both the RealEstate10K and our *OmniWorld-Game* benchmarks. This result validates *OmniWorld* as a powerful training resource for enhancing a model's capability to follow complex camera instructions in dynamic environments.

5 Related Work

World model dataset. The ability of models to perform world modeling is intrinsically linked to the availability of large-scale, high-quality spatio-temporal datasets. Static 3D datasets (Dai et al., 2017; Silberman et al., 2012; Li & Snavely, 2018) have advanced 3D reconstruction by providing precise geometric information. However, their static nature limits their utility for modeling motion. In video generation, large-scale video-text datasets (Chen et al., 2024; Bain et al., 2021; Nan et al., 2024; Ju et al., 2024) offer rich semantic annotations but lack geometric information (e.g., depth), making them unsuitable for 4D world modeling. To bridge this gap, researchers have created dynamic real-world datasets for autonomous driving (Geiger et al., 2013; Sun et al., 2020) and human-robot interaction (Palazzolo et al., 2019; Liu et al., 2022; Damen et al., 2018; Fang et al., 2024). While valuable, these datasets often suffer from a lack of scene diversity and noisy geometric annotations. With advancements in modern rendering technology significantly reducing the sim-to-real gap (Wang et al., 2020), synthetic datasets have emerged as a valuable alternative providing precise annotations. However, recent synthetic datasets (Butler et al., 2012; Mayer et al., 2016; Harley et al., 2022; Wang et al., 2020; Karaev et al., 2023; Mehl et al., 2023) still fall short in terms of scale, diversity, and modal richness compared to our *OmniWorld-Game* dataset (Tab. 1).

Table 6: **Comparison of original and fine-tuned models for video depth estimation** on Sintel (Butler et al., 2012), Bonn (Palazzolo et al., 2019) and KITTI (Geiger et al., 2013). * denotes models that have been fine-tuned on *OmniWorld*.

Method	Align	Sintel		Bonn		KITTI	
		Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
DUSt3R (Wang et al., 2024c)	scale	0.652	0.436	0.151	0.839	0.143	**0.814**
DUSt3R*		**0.512**	**0.456**	**0.083**	**0.920**	**0.135**	0.800
CUT3R (Wang et al., 2025b)	scale	0.417	0.510	0.078	0.937	0.123	0.875
CUT3R*		**0.396**	**0.516**	**0.078**	**0.938**	**0.107**	**0.907**
DUSt3R (Wang et al., 2024c)	scale&shift	0.570	**0.493**	0.152	0.835	**0.135**	**0.818**
DUSt3R*		**0.520**	0.480	**0.084**	**0.914**	0.136	0.808
CUT3R (Wang et al., 2025b)	scale&shift	0.537	0.556	0.075	0.944	0.111	0.884
CUT3R*		**0.314**	**0.574**	**0.067**	**0.964**	**0.103**	**0.912**

Table 7: **Comparison of original and fine-tuned models for camera-controlled video generation evaluation** on RealEstate10K (Zhou et al., 2018) and *OmniWorld-Game* benchmark. * denotes models that have been fine-tuned on *OmniWorld*.

Method	Benchmark	TransErr↓	RotErr↓	CamMC↓	FVD	
					VideoGPT↓	StyleGAN↓
AC3D (Bahmani et al., 2024)	RealEstate10K	3.4433	0.6308	3.6615	479.320	**409.795**
AC3D*		**2.8648**	**0.5314**	**3.0518**	**472.683**	416.948
AC3D (Bahmani et al., 2024)	*OmniWorld-Game*	6.2788	0.8867	6.6965	1745.778	1594.885
AC3D*		**4.1428**	**0.7610**	**4.4854**	**1437.247**	**1249.186**

3D geometric foundation models. 3D geometric foundation models have recently emerged as a data-driven alternative to traditional methods. Early works like DUSt3R (Wang et al., 2024c) and MonST3R (Zhang et al., 2024) operate on image pairs, requiring expensive global alignment for larger scenes. Further research has introduced diverse architectural innovations to overcome this, including parallel processing (Fast3R (Yang et al., 2025a)), decomposing the learning task (FLARE (Zhang et al., 2025)), online processing for image streams (CUT3R (Wang et al., 2025b)), multi-task learning (VGGT (Wang et al., 2025a)), and permutation-equivariant designs (π^3 (Wang et al., 2025d)). However, the efficacy of these models is fundamentally tied to large-scale, multi-modal training data. We validate *OmniWorld* as a powerful training resource that fulfills this need.

Camera-controlled video generation. Most methods in this field inject camera parameters (such as Plücker embeddings) into a pre-trained video diffusion model (Blattmann et al., 2023; Chen et al., 2023; Yang et al., 2024b) with representative works including MotionCtrl (Wang et al., 2024d), CameraCtrl (He et al., 2024), CAMI2V (Zheng et al., 2024), AC3D (Bahmani et al., 2024). Despite this progress, these methods still struggle to generate dynamic content with complex camera control. They are typically trained on datasets like RealEstate10K (Zhou et al., 2018) or DL3DV-10K (Ling et al., 2024), which consist of static scenes with smooth camera motions. This data limitation inherently restricts them to handle dynamic scenes (He et al., 2025a). The performance gap is evident on our challenging *OmniWorld-Game* benchmark.

6 Conclusion

We introduce *OmniWorld*, a large-scale, multi-domain, and multi-modal dataset designed to address the critical data bottleneck for 4D world modeling. By integrating self-collected *OmniWorld-Game* dataset and several public datasets from various domains, we create a comprehensive data resource for 4D world modeling. We demonstrate that *OmniWorld-Game* serves as a challenging benchmark for 3D geometric prediction and camera-controlled video generation, revealing the limitations of current methods. Furthermore, we provide strong evidence that fine-tuning with *OmniWorld* significantly boosts the performance of these models, underscoring its value as a powerful training resource. We believe that *OmniWorld* will serve as a crucial data resource for the community, accelerating the development of more general and robust models for understanding and interacting with the real physical world.

ACKNOWLEDGEMENTS

This work is supported by Shanghai Artificial Intelligence Laboratory.

ETHICS STATEMENT

Our work, the *OmniWorld* dataset, is a composite dataset consisting of a newly collected game-derived dataset (*OmniWorld-Game*) and several curated public datasets. We have undertaken a multi-faceted approach to ensure our practices for both collecting new data and curating existing data are legally compliant.

1) Responsible Acquisition of Game-Derived Data. Game data was captured from legally purchased games using standard, non-invasive tools (e.g., OBS, ReShade) without any reverse engineering or cheating. To respect the source material, we automatically remove UI elements and text, and manually filter for sensitive content such as story spoilers.

2) Adherence to Terms of Use for Game Content. Our use of game content is strictly non-commercial, aligning with publisher terms of service (e.g., Rockstar Games (Rockstar Games, 2024)). The dataset is intended solely to advance academic research and does not compete with or infringe upon the economic interests of the copyright holders.

3) Curation of Public Datasets. *OmniWorld* also incorporates public datasets to enhance domain diversity. We have strictly adhered to the original license of each dataset, ensuring proper attribution and compliance with all usage terms.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, 2012. URL https://api.semanticscholar.org/CorpusID:4637111.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7063–7072, 2019.

OBS Contributors. Obs studio, 2024. URL `https://obsproject.com/`.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Google DeepMind. Genie 3. `https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/`, 2025. Accessed: 2025-08-27.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, 2018.

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. *arXiv preprint arXiv:2412.08376*, 2024.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control, 2024. URL `https://arxiv.org/abs/2412.03568`.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

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

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Annika Hagemann, Moritz Knorr, and Christoph Stiller. Deep geometry-aware camera self-calibration from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3438–3448, 2023.

Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pp. 59–75. Springer, 2022.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3406–3416, June 2025.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions, 2024. URL https://arxiv.org/abs/2407.06358.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Hoang-An Le, Partha Das, Thomas Mensink, Sezer Karaoglu, and Theo Gevers. EDEN: Multi-modal Synthetic Dataset of Enclosed garDEN Scenes. In *Proceedings of the IEEE/CVF Winter Conference of Applications on Computer Vision (WACV)*, 2021.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *arxiv*, 2024.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21013–21022, June 2022.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17810–17820, 2025.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. *arXiv*, 2019. URL `https://arxiv.org/abs/1905.02082`.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL `https://arxiv.org/abs/2408.00714`.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

ReShade Contributors. ReShade, 2024. URL `https://reshade.me/`.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pp. 102–118. Springer, 2016.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Rockstar Games. Policy on Posting Copyrighted Rockstar Games Material, 2024. URL https://support.rockstargames.com/articles/7bNaeoMFTV0iUDGhStTXvz/policy-on-posting-copyrighted-rockstar-games-material.

Chris Rockwell, Joseph Tung, Tsung-Yi Lin, Ming-Yu Liu, David F Fouhey, and Chen-Hsuan Lin. Dynamic camera poses and where to find them. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12444–12455, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *CVPR*, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *arXiv preprint arXiv:2412.06974*, 2024.

Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025a.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025b.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025b.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024b. URL https://arxiv.org/abs/2410.19115.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025c. URL https://arxiv.org/abs/2507.02546.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024c.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20270–20281, October 2023.

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, 2025d. URL https://arxiv.org/abs/2507.13347.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025e. URL https://arxiv.org/abs/2505.10565.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024d.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *CVPR*, 2025.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22378–22389, 2024.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024a.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025a.

Rujia Yang, Geng Chen, Chuan Wen, and Yang Gao. Fp3: A 3d foundation policy for robotic manipulation. *arXiv preprint arXiv:2503.08950*, 2025b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 325–334, 2011.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos, 2025b.

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025.

Chengbo Yuan, Suraj Joshi, Shaoting Zhu, Hang Su, Hang Zhao, and Yang Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2503.18738*, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views, 2025. URL `https://arxiv.org/abs/2502.12138`.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

A OVERVIEW

This appendix provides comprehensive supplementary materials, organized as follows:

- **Dataset & Benchmarks:** Sec. B details the *OmniWorld* dataset construction. Extended details on our benchmark and fine-tuning experiments are provided in Sec. C and Sec. D, supplemented by comprehensive benchmark statistics in Sec. I.
- **Data Processing & Validation:** Sec. F details our strategy for filtering high-quality video clips. Sec. G and Sec. J provide qualitative and quantitative validations of our annotation pipelines, respectively.
- **Analysis:** Sec. H presents a comprehensive failure case analysis to discuss the current limitations.

B *OmniWorld* DATASET

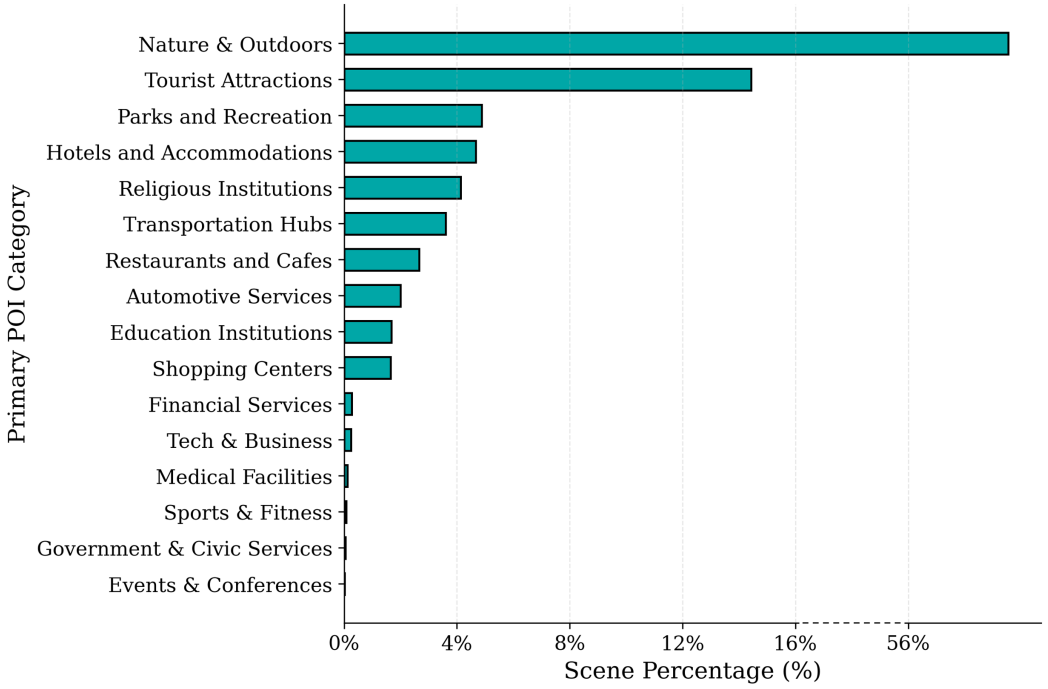


Figure 6: **The *OmniWorld-Game* distribution** of scene category (the primary POI locations).

To quantitatively analyze the scene diversity of *OmniWorld-Game*, we adopt the methodology from DL3DV (Ling et al., 2024) to classify and count scenes across 16 Point-of-Interest (POI) categories (Ye et al., 2011). The statistical results are shown in Fig. 6. *OmniWorld-Game* encompasses a wide variety of scene categories, including "Nature & Outdoors", "Tourist Attractions", "Parks and Recreation", and "Hotels and Accommodations". "Nature & Outdoors" represents the largest share, reflecting its dominant presence in the dataset. The distribution of these scene categories aligns with their prevalence in the real world and the characteristics of the games themselves. For instance, scenes related to "Government & Civic Services" and "Events & Conferences" are typically less frequent in games, leading to their lower representation in our dataset. These statistics further validate the richness and real-world attributes of *OmniWorld-Game*.

To provide a more detailed analysis of the dominant "Nature & Outdoors" scenes in *OmniWorld-Game*, we further subdivide this category into 5 second-level and 40 third-level categories. The detailed distribution is shown in Fig. 7. Our statistics reveal that "Natural Landforms & Ecosystems" is the dominant second-level category. Within this category, scenes depicting "Forests &

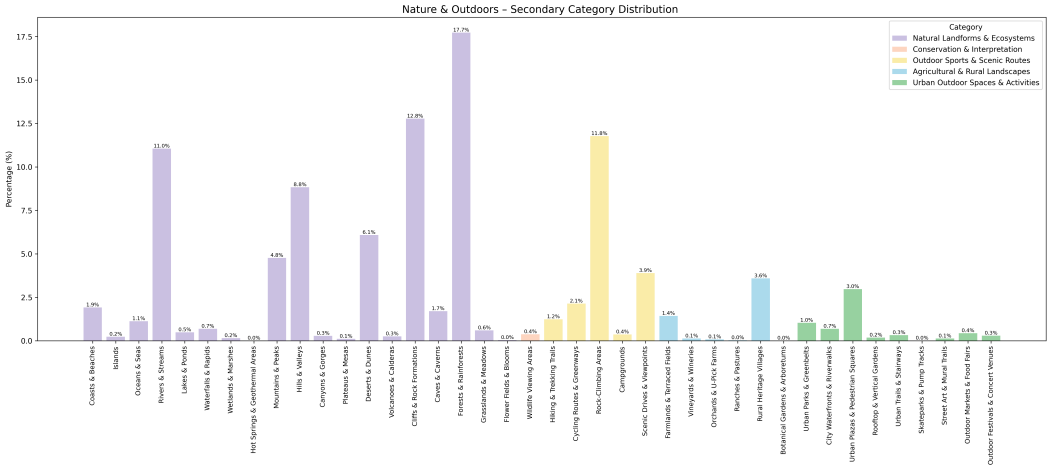


Figure 7: **Scene diversity within the "Nature & Outdoors" category.** A quantitative breakdown of second- and third-level scene categories in *OmniWorld-Game* dataset, demonstrating the high internal diversity and distribution of natural environments.

Rainforests" and "Cliffs & Rock Formations" are the most prevalent. "Outdoor Sports & Scenic Routes" is the second-largest category, with scenes of "Rock-Climbing Areas" and "Scenic Drives & Viewpoints" being particularly prominent. Additionally, "Urban Outdoor Spaces & Activities" and "Agricultural & Rural Landscapes" also make up a small portion of the data. These detailed statistics confirm that the "Nature & Outdoors" scenes in *OmniWorld-Game* are not only abundant but also internally diverse. This rich composition provides a diverse data source for world modeling in complex natural environments.

C *OmniWorld-Game* BENCHMARK

C.1 3D GEOMETRIC PREDICTION

Experiment details. We adhere to the default configurations of each evaluated model. The entire evaluation process is conducted on a single A800 GPU. All images are consistently resized to a long side of 512 pixels while preserving aspect ratio.

For the monocular depth estimation, we evaluate the first 200 frames of 18 test sequences from the *OmniWorld-Game* benchmark. Following the evaluation protocols of prior works (Zhang et al., 2024; Wang et al., 2025b;d), we focus on scale-invariant monocular depth accuracy. The primary evaluation metrics are Absolute Relative Error (Abs Rel) and threshold accuracy ($\delta < 1.25$). Under this setting, the depth map of each frame is independently aligned with its corresponding ground truth.

For the video depth estimation, we select the first 100 frames of the same test sequence from the *OmniWorld-Game* benchmark. To ensure a fair comparison across all models, we cap the input sequence length at 100 frames, as some models (e.g., FLARE (Zhang et al., 2025)) cannot handle longer sequences without errors. Similar to the monocular depth estimation, we report Abs Rel and $\delta < 1.25$. To more comprehensively evaluate depth consistency across video sequences, we provide results under two different alignment settings: (i) scale-only alignment (scale) and (ii) combined scale and translation alignment (scale & shift). These settings test a model's depth estimation capabilities under different constraints, particularly in handling motion and viewpoint changes.

It is important to note that since the benchmark data is included in the training set of π^3 (Wang et al., 2025d), we did not evaluate it in our benchmark.

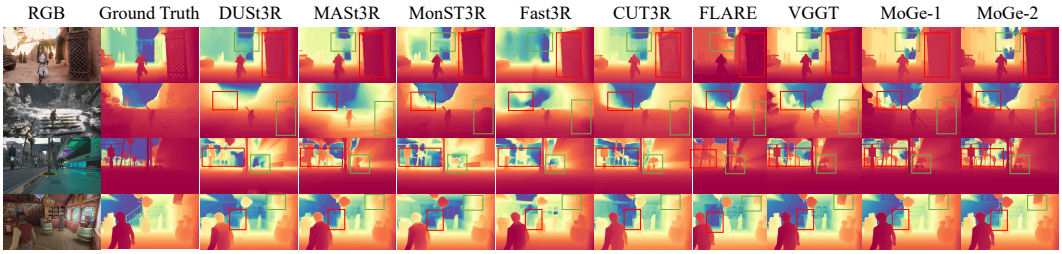


Figure 8: **Qualitative comparison of monocular depth estimation** on *OmniWorld-Game* benchmark across various methods.

Table 8: **Comparison of original and fine-tuned models for camera pose estimation** on Sintel (Butler et al., 2012), TUM-dynamics (Sturm et al., 2012) and ScanNet (Dai et al., 2017). * denotes models that have been fine-tuned on *OmniWorld*.

Method	Sintel			TUM-dynamics			ScanNet		
	ATE↓	RPE trans↓	RPE rot↓	ATE↓	RPE trans↓	RPE rot↓	ATE↓	RPE trans↓	RPE rot↓
CUT3R (Wang et al., 2025b)	0.210	0.071	**0.627**	0.045	0.014	0.441	0.096	0.022	0.733
CUT3R*	**0.178**	**0.055**	0.651	**0.041**	**0.013**	**0.374**	**0.095**	**0.022**	**0.604**

C.2 Camera-Controlled Video Generation

Experiment details. AC3D (Bahmani et al., 2024) uses CogVideoX-5B (Yang et al., 2024b) as base T2V model, it generates 25 frames per inference at a resolution of 480 × 720. CamCtrl (He et al., 2024) and MotionCtrl (Wang et al., 2024d) use Stable Video Diffusion (SVD) (Blattmann et al., 2023) as base I2V model and generate 14-frame video sequences at a resolution of 320 × 512. CAMI2V (Zheng et al., 2024) uses DynamiCrafter (Xing et al., 2023) as base I2V model. It generates 16-frame video sequences at a resolution of 320 × 512. For a fair comparison with CamCtrl and MotionCtrl, we use the first 14 frames of its generated videos for evaluation. We use π^3 (Wang et al., 2025d) to get camera poses of the generated videos. All methods are evaluated on an A800 GPU.

C.3 Additional Qualitative Results

We provide additional qualitative results that complement the quantitative analysis in Sec. 3.1 of the main paper. Fig. 8 provides a visual comparison of monocular depth estimation results from various methods on our *OmniWorld-Game* benchmark. These visualizations confirm that MoGe-2 (Wang et al., 2025c) generates depth maps with significantly sharper details and more coherent geometric structures compared to its counterparts.

D Model Fine-tuning

D.1 Camera Pose Estimation.

Following prior work (Wang et al., 2025b;d), we report the Absolute Trajectory Error (ATE), Relative Pose Error for translation (RPE trans), and Relative Pose Error for rotation (RPE rot) on Sintel (Butler et al., 2012), TUM-dynamics (Sturm et al., 2012) and ScanNet (Dai et al., 2017). The results in Tab. 8 show that CUT3R's performance notably improved after fine-tuning on *OmniWorld* in camera pose estimation.

We perform relative camera pose evaluation on the DynPose-100K (Rockwell et al., 2025) and the *OmniWorld-CityWalk* test set. Following prior work (Dong et al., 2024), we assess performance with three indicators: AUC@5/10/20, which measure the area under the pose accuracy curve. This curve is based on minimum thresholds of 5, 10, and 20 degrees for rotation and translation angular errors. Reloc3r (Dong et al., 2024) demonstrated substantial improvements in its ability to estimate dynamic camera poses after fine-tuning on *OmniWorld* in relative camera pose evaluation (Tab. 9).

Table 9: **Comparison of original and fine-tuned models for relative camera pose evaluation** on DynPose-100K (Rockwell et al., 2025), *OmniWorld-CityWalk* (Li et al., 2025). * denotes models that have been fine-tuned on *OmniWorld*.

Method	DynPose-100K			*OmniWorld-CityWalk*		
	AUC@5↑	AUC@10↑	AUC@20↑	AUC@5↑	AUC@10↑	AUC@20↑
Reloc3r (Dong et al., 2024)	6.9	15.4	27.1	33.3	49.4	63.1
Reloc3r*	**14.4**	**25.5**	**37.8**	**42.5**	**58.0**	**70.3**

D.2 Implementation Details

We conduct comprehensive fine-tuning experiments on several SOTAs to validate the efficacy of our *OmniWorld* as a training resource. All experiments are performed on 8 A800 GPUs.

DUSt3R (Wang et al., 2024c). For fine-tuning, we use *OmniWorld-Game* alongside a portion of DUSt3R's original training sets, including ARKitScenes (Baruch et al., 2021), MegaDepth (Li & Snavely, 2018), and Waymo (Sun et al., 2020). We load the pre-trained weights of DUSt3R and performed full fine-tuning. The model is fine-tuned on images with random resolutions (e.g., 288×512, 384×512, 336×512). The training runs for 40 epochs, with each epoch consisting of 800 iterations. We use the AdamW optimizer with an initial learning rate of 2.5×10^{-5} and a weight decay of 0.05. Each GPU had a batch size of 7, with each batch containing two images.

CUT3R (Wang et al., 2025b). We fine-tune CUT3R using *OmniWorld-Game* and a subset of its original training data, including CO3Dv2 (Reizenstein et al., 2021), WildRGBD (Xia et al., 2024), ARKitScenes (Baruch et al., 2021), Waymo (Sun et al., 2020), and TartanAir (Wang et al., 2020). We load the pre-trained weights and follow the training strategy from CUT3R's training stage 3. We fine-tune on higher-resolution images with varied aspect ratios, setting the maximum side to 512 pixels. The encoder is frozen, with only the decoder and heads being trained on longer sequences of 4 to 64 views. The model is fine-tuned for 2,000 iterations with a total batch size of 96 and a learning rate of 1.0×10^{-6}, optimized by AdamW with a weight decay of 0.05.

Reloc3r (Dong et al., 2024). For fine-tuning Reloc3r, we utilize *OmniWorld-Game*, *OmniWorld-CityWalk*, *OmniWorld-HoloAssist*, and *OmniWorld-EpicKitchens*, along with a portion of its original training sets, including CO3Dv2 (Reizenstein et al., 2021), ARKitScenes (Baruch et al., 2021), Scannet++ (Yeshwanth et al., 2023), BlendedMVS (Yao et al., 2020), and MegaDepth (Li & Snavely, 2018). We load the pre-trained weights, freeze the ViT encoder, and only update the weights for the decoder and pose regression head. Fine-tuning is performed on images of random resolutions, including 288 $\times$ 512, 384 $\times$ 512, and 336 $\times$ 512. The model is trained for 80 epochs, with each epoch comprising 400 iterations. We use the AdamW optimizer with a learning rate of 5.0×10^{-6} and a weight decay of 0.05. Each GPU has a batch size of 32, with each batch containing two images.

AC3D (Bahmani et al., 2024). We fine-tune AC3D using *OmniWorld-Game*, *OmniWorld-EpicKitchens*, *OmniWorld-HOI4D*, *OmniWorld-HoloAssist*, *OmniWorld-EgoExo4D*, and *OmniWorld-EgoDex*, as well as the original training set, RealEstate10K (Zhou et al., 2018). We load the pre-trained weights of the AC3D ControlNet (Zhang et al., 2023), which is based on CogVideoX-5B (Yang et al., 2024b). Only the ControlNet model is fine-tuned, with other network structures frozen. The fine-tuning is performed on video clips of 49 frames with a resolution of 352 $\times$ 640. The model is fine-tuned for 6,000 iterations with a total batch size of 8 and a learning rate of 5.0×10^{-5}, optimized by AdamW with a weight decay of 0.0001. The fine-tuned and original models are evaluated on two distinct benchmarks: a random subset of 150 video samples from the RealEstate10K (Zhou et al., 2018) test set and *OmniWorld-Game* benchmark, which consists of 200 video samples. For a fair comparison, all models are configured to output videos at a uniform resolution of 720 $\times$ 480 with a sequence length of 25 frames.

D.3 Visual Results.

Fig. 9 provides a qualitative comparison of DUSt3R (Wang et al., 2024c) and CUT3R (Wang et al., 2025b) on the Sintel (Butler et al., 2012) subset of the Video Depth Estimation benchmark, evaluated both before and after fine-tuning on *OmniWorld*. After fine-tuning, both models recover finer

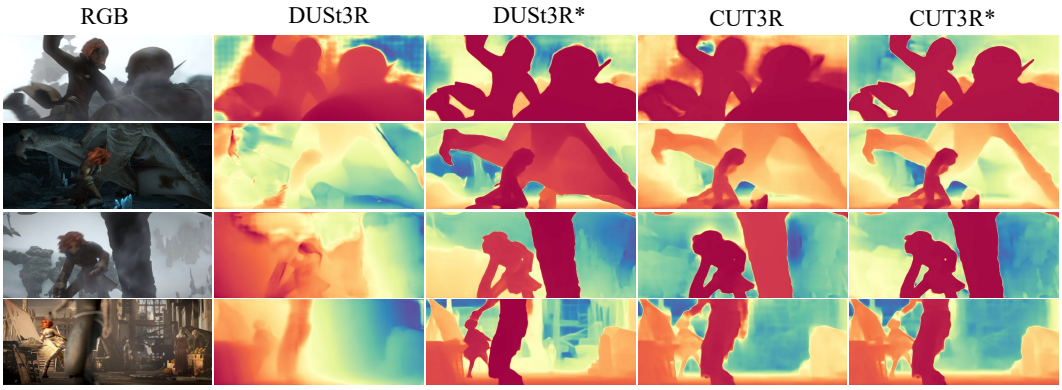


Figure 9: **Qualitative comparison of original and fine-tuned models for video depth estimation** on the Sintel (Butler et al., 2012). * denotes models that have been fine-tuned on *OmniWorld*. After fine-tuning, both models recover finer geometric details and produce more accurate depth maps, highlighting the efficacy of *OmniWorld* as a geometric supervision source.

Figure 10: **Qualitative comparison of original and fine-tuned models for camera-controlled video generation.** * denotes models that have been fine-tuned on *OmniWorld*. The visualizations show that fine-tuning with our dataset significantly improves the model's ability to generate videos that more accurately follow camera trajectories and maintain higher temporal consistency for moving objects.

geometric details and generate more accurate depth maps. These results indicate that *OmniWorld* offers strong geometric supervision and can substantially enhance a model's geometric prediction capability.

Fig. 10 presents a visual comparison of AC3D (Bahmani et al., 2024) on the *OmniWorld-Game* benchmark before and after fine-tuning on the *OmniWorld* dataset for the camera-controlled video generation task. The visualizations clearly show that after fine-tuning, the generated videos more closely follow the desired camera trajectory and exhibit higher temporal consistency for moving objects. This demonstrates that *OmniWorld* can significantly enhance a model's ability to model dynamics.

E Statement on LLM Usage

In the preparation of this manuscript, we use Large Language Models (LLMs) only to polish writing.

Figure 11: **Examples of low-quality frames discarded by the filtering pipeline.** Each column illustrates a specific failure case: Texture Scarcity, Severe Underexposure, Dynamic Occlusion, Motion Blur.

F HIGH-QUALITY VIDEO CLIP SELECTION

To ensure the precision and geometric consistency of camera pose annotations within the *OmniWorld* dataset, we implemented a sophisticated, multi-stage automated filtering pipeline. This pipeline is designed to extract high-quality, texture-rich, and smoothly moving video segments (Splits) from raw, long video sequences, while discarding frames unsuitable for robust 3D reconstruction. The filtering process encompasses two primary stages: frame-level quality assessment and temporal consistency validation.

F.1 FRAME-LEVEL QUALITY ASSESSMENT

Before sequence segmentation, each individual frame undergoes a rigorous evaluation. Any frame failing to meet the following criteria is designated as an "invalid frame," triggering a truncation of the current video segment.

Texture Scarcity. Robust feature matching is fundamental for accurate camera pose estimation. We employ the SIFT (Lowe, 2004) algorithm for feature point extraction. A frame is discarded if it yields few valid keypoints, indicative of homogenous surfaces or extreme blur.

Geometric Invalidity. For source data including depth information, we check the integrity of the depth maps. A frame is deemed geometrically invalid if it contains invalid values or if the area of invalid depth (zero values) exceeds 60% of the total image pixels.

Severe Underexposure. Extremely dark scenes significantly degrade the signal-to-noise ratio, adversely affecting reconstruction quality. We quantify the proportion of pixels with values below 20 in the RGB image. If dark pixels constitute over a specific threshold of the frame, it is marked as unusable.

Dynamic Occlusion. Large-area dynamic foreground objects can confound SfM algorithms, which typically assume a static scene. Leveraging semantic masks generated by Grounding DINO (Liu et al., 2023) and SAM2 (Ravi et al., 2024), we calculate the screen occupancy of dynamic entities (e.g., characters, vehicles). If the dynamic region exceeds 60% of the frame, it is consequently excluded.

F.2 TEMPORAL CONSISTENCY AND MOTION FILTERING

To guarantee the continuity of inter-frame relationships within video segments, we integrate optical flow estimation, utilizing RAFT (Teed & Deng, 2020), to impose constraints on adjacent frame motion characteristics.

Forward-Backward Flow Consistency. We compute bidirectional optical flow between frames at time t and $t + 1$, and employ the forward-backward error to detect occlusions and matching inaccuracies. If the proportion of pixels satisfying geometric consistency constraints falls below 50%, it indicates a sudden scene change or severe occlusion, leading to the truncation of the current sequence.

Motion Magnitude Limitation. Excessive camera motion often results in motion blur and reduced inter-frame overlap, which can severely impede feature tracking. We calculate the mean magnitude of normalized optical flow. If the average motion exceeds 10% of the image dimensions (a threshold set at 0.1), it is classified as rapid motion, and the current segment is truncated. This criterion is crucial for preventing tracking loss during SfM.

Fig. 11 provides visual examples of various "bad cases" that our filtering pipeline effectively identifies and discards. These examples illustrate frames affected by texture scarcity, severe underexposure, dynamic occlusion, and significant motion blur, all of which compromise the quality of camera pose estimation.

Through this stringent filtering methodology, we effectively eliminate low-quality data, thereby ensuring that each selected video segment within *OmniWorld*is suitable for generating high-precision camera pose annotations.

G Qualitative Validation of Annotation Pipelines

To visually substantiate the effectiveness of our annotation pipelines, we provide qualitative validation.

G.1 Camera Pose Estimation

We compare our two-stage pipeline (DroidCalib (Hagemann et al., 2023) initialized, refined via CoTracker3 (Karaev et al., 2024) and Bundle Adjustment) against the baseline DroidCalib (Hagemann et al., 2023). To visualize the accuracy of the estimated camera poses, we project the ground truth depth maps into a unified 3D point cloud using the estimated camera extrinsic parameters. As shown in Fig. 12, the baseline DroidCalib often suffers from drift or misalignment in dynamic scenes, resulting in "ghosting" artifacts or structural inconsistencies (highlighted by red arrows). In contrast, our pipeline produces globally consistent point clouds where static background geometries (e.g., walls, pillars, trees) are sharply aligned, demonstrating superior pose accuracy in the presence of dynamic foregrounds.

G.2 Foreground Mask Generation in Dynamic Scenarios

We compare our semantic segmentation pipeline (Grounding DINO (Liu et al., 2023) with SAM (Ravi et al., 2024)) against the recent SOTA video-specific segmentation model SegAnyMo (Huang et al., 2025).

As illustrated in Fig. 13, our pipeline demonstrates superior mask results and robustness in complex dynamic environments. In the second row (narrow alley), our method successfully segments multiple walking pedestrians that are missed by the baseline. In the third row (road scene), SegAnyMo incorrectly segments static traffic signs as dynamic foregrounds, whereas our method correctly identifies and segments the moving vehicles. These results validate our choice of the Grounding DINO with SAM pipeline for generating reliable foreground masks, which are crucial for filtering dynamic interference in downstream tasks.

H Failure Case Analysis

While our automated pipeline (Grounding DINO (Liu et al., 2023) with SAM (Ravi et al., 2024)) demonstrates high reliability across diverse domains, we provide a transparent analysis of typical failure cases to highlight potential limitations. We identify two primary scenarios where the pipeline may exhibit imperfections: extreme crowd density and color ambiguity in close-range interactions.

As shown in the top row of Fig. 14, in highly cluttered outdoor scenes with numerous dynamic agents, the pipeline successfully segments the majority of pedestrians. However, due to resolution limitations and severe occlusion, detection may fail for distal subjects (i.e., small figures in the far background).

In ego-centric robotic manipulation scenarios (bottom row of Fig. 14), the camera often operates in close proximity to objects. When the robotic arm interacts with objects that share similar color

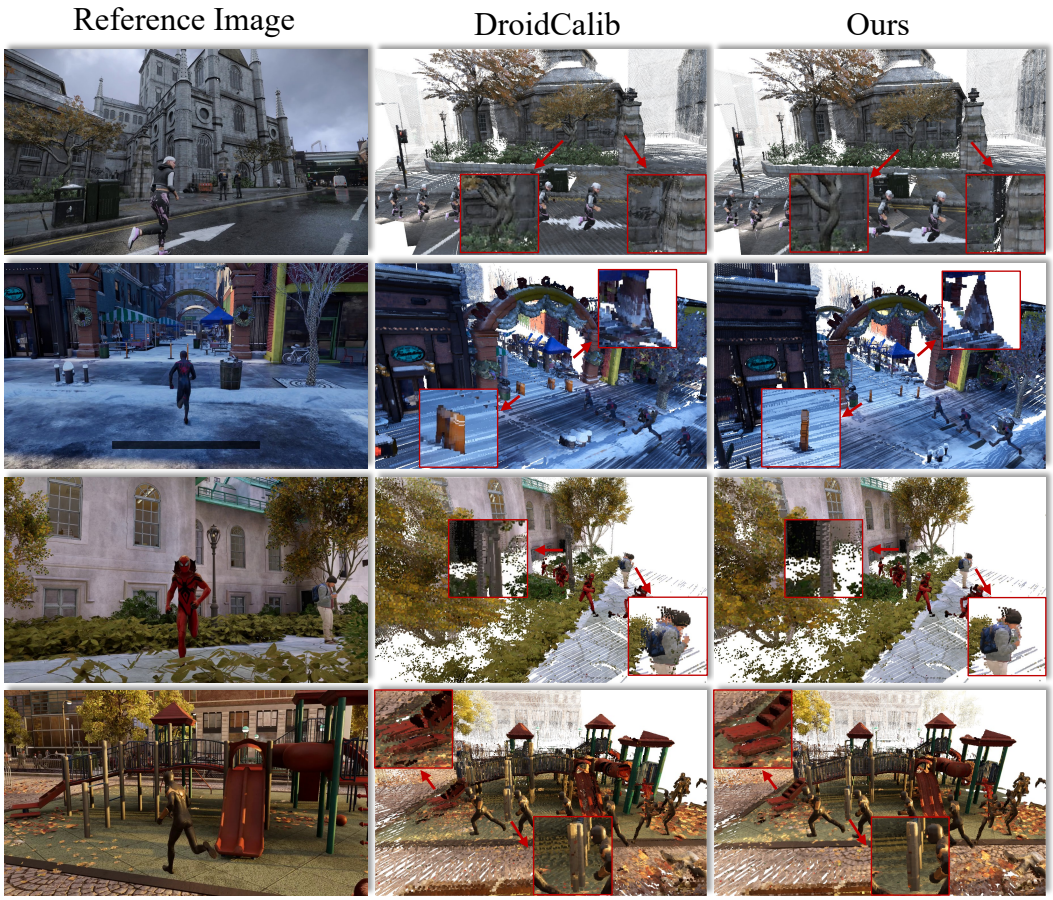


Figure 12: **Qualitative comparison of point cloud reconstructions between DroidCalib (Hagemann et al., 2023) and our camera pose estimation pipeline.** We visualize the consistency of camera poses by accumulating point clouds over a video sequence. The left column shows the reference frame. The middle column shows the reconstruction using poses from the baseline DroidCalib, where red arrows indicate significant misalignment and ghosting artifacts on static structures. The right column shows the reconstruction using ours, which effectively resolves these artifacts, resulting in sharper and more geometrically consistent 3D structures.

textures or are spatially adjacent, the segmentation mask may exhibit semantic leakage, inadvertently covering nearby static objects along with the moving arm.

Crucially, empirical observations suggest that these localized annotation imperfections have a negligible impact on our primary downstream task: dynamic camera pose estimation. As long as the dominant dynamic foregrounds are masked and a sufficient portion of the static background remains visible, the pose estimation remains accurate and stable.

I Detailed Benchmark Statistics and Analysis

I.1 Geometric Prediction Benchmark

To ensure a comprehensive evaluation of SOTA models, we employed an "Attribute-Balanced Sampling" strategy for the *OmniWorld-Game* benchmark. This section details the quantitative distributions and qualitative definitions of our test samples.

We curated a set of 90 long-sequence samples, each containing 384 frames, specifically designed to evaluate long-term geometric consistency. Each sample lasts for 16 seconds.

Figure 13: **Qualitative comparison of foreground mask generation on in-the-wild videos.** We compare SegAnyMo (Huang et al., 2025) with our pipeline. SegAnyMo struggles to consistently track and segment all moving subjects, while our method provides precise masks for the moving subjects.

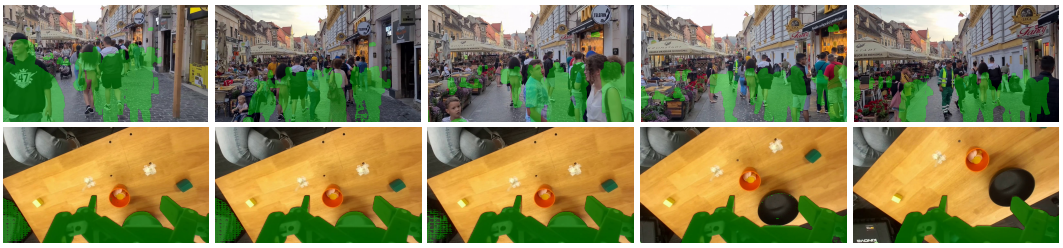


Figure 14: **Failure case analysis of foreground mask annotations.** We visualize the segmentation masks (green overlay) in challenging scenarios.

Scene Distribution. As illustrated in Fig. 15a, the dataset maintains a balance across Outdoor-Urban (58%), Outdoor-Natural (18%), and Indoor (24%) environments. This distribution ensures that models are rigorously tested on both unbounded scenes with complex backgrounds and bounded scenes featuring intricate internal structures.

Dynamic Complexity. We categorize motion into three distinct levels based on the intensity of object and camera movement. High-Dynamic (48%): Features intense motions such as running, flying, or rapid vehicle movement. Medium-Dynamic (38%): Includes standard motion (walking, running) and regular interactions. Low-Dynamic (14%): Primarily consists of background environmental motion or still character animations.

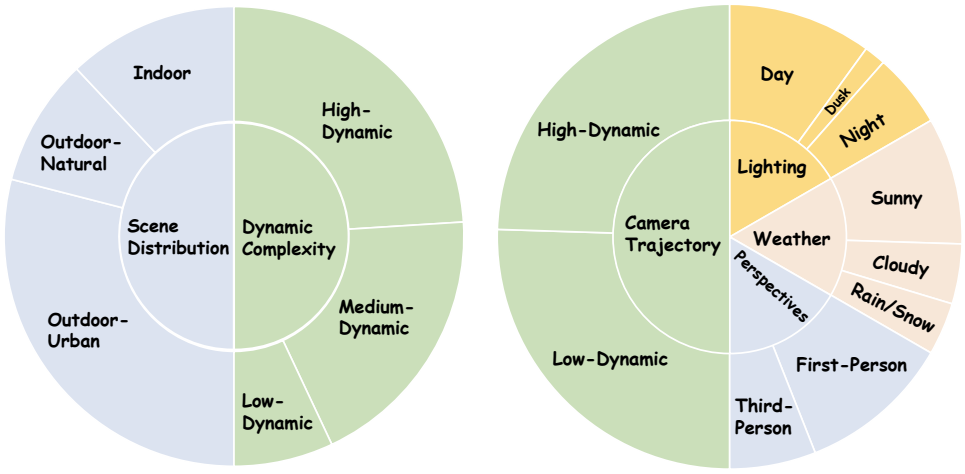
I.2 Camera-Controlled Video Generation Benchmark

For the camera-controlled video generation task, we selected 200 samples, prioritizing trajectory complexity and environmental richness to ensure a challenging evaluation.

Camera Trajectory Complexity. As depicted in Fig. 15b, we split the samples based on camera motion. High-Complexity (51%): Features rapid translations combined with rotations, or compound high-speed movements. Medium-Complexity (49%): Represents distinct but stable motions, typical of cinematic tracking shots.

Environmental Diversity. The benchmark spans a wide range of lighting conditions (Day 60%, Night 31%, Dusk 9%) and weather scenarios (Sunny 53%, Cloudy 25%, Rain/Snow 22%).

Perspective. We include both First-person (64%) and Third-person (36%) views to test generation robustness across different field-of-view dynamics.



(a) Geometric Prediction Benchmark (b) Video Generation Benchmark

Figure 15: **Statistical distribution of the *OmniWorld-Game* benchmark samples.** (a) The geometric prediction benchmark is balanced across scene types (Indoor, Outdoor) and dynamic complexity levels. (b) The video generation benchmark covers a diverse range of camera trajectory complexities, environmental conditions, and perspectives.

As shown in the visual examples in Fig. 16, *OmniWorld-Game* high-dynamic sequences introduce complex camera trajectories and drastic geometric changes, posing a substantial challenge for both geometric prediction and camera-controlled video generation.

J Quantitative Validation of Annotation Pipelines

To ensure the trustworthiness of the *OmniWorld* dataset, we conducted extensive quantitative analyses in key modalities.

J.1 Depth Annotation Validation

The accuracy of our ground truth (GT) depth in *OmniWorld-Game* is intrinsically guaranteed by the rendering engine and validated by the fine-tuning experiments in Sec. 4.1. Here, we focus on validating the quality of our pseudo-labeled depth for the DROID (Khazatsky et al., 2024) dataset, which was generated using FoundationStereo (Wen et al., 2025).

To assess the utility of these annotations in downstream tasks, we pre-trained the FP3 (Yang et al., 2025b) model on point clouds projected from two sources: (1) the original DROID depth, and (2) our refined depth annotations. We then evaluated these models on four real-world tasks.

As presented in Tab. 10, the model pre-trained on our annotated depth yields significantly higher success rates across all tasks compared to the baseline using original DROID depth. This result demonstrates that our depth annotations preserve better geometric consistency.

Table 10: **Real-world manipulation task success rates (%).** Comparison of the FP3 pre-trained on the original DROID depth and our refined depth annotations. Our annotations consistently lead to higher success rates.

Method	Open Drawer	Stack Cups	Pick up Toy	Put Toy into Basket
FP3 (Original DROID depth)	25	10	60	35
FP3 (Ours refined depth)	**40**	**35**	**90**	**55**

J.2 CAMERA POSE ANNOTATION VALIDATION

We evaluate our camera pose annotation pipeline in two scenarios: datasets without GT depth (e.g., in-the-wild videos) and datasets with GT depth (e.g., *OmniWorld-Game*).

Scenario 1: Evaluation on Data without GT Depth. We compared our full pipeline (VGGT (Wang et al., 2025a) initialization followed by CoTracker3 (Karaev et al., 2024) and Bundle Adjustment) against the baseline VGGT on the Sintel benchmark (Butler et al., 2012). As shown in Tab. 11, our method significantly outperforms the baseline, reducing the Absolute Trajectory Error (ATE) by over 50%. This confirms that our optimization strategy effectively refines coarse initializations into precise trajectories.

Table 11: **Pose estimation performance on the Sintel benchmark.** Comparison between the baseline VGGT and our full annotation pipeline.

Method	ATE ↓	RPE trans ↓	RPE rot ↓
VGGT (Baseline)	0.167	0.062	0.491
Ours	**0.082**	**0.042**	**0.246**

Scenario 2: Evaluation on Data with GT Depth. To evaluate the reliability of our camera pose annotations where GT depth is available, we adopted the rigorous validation protocol following (Rockwell et al., 2025). We conducted a large-scale evaluation across 8,345 randomly sampled frame pairs from our dataset.

For each pair, we extracted high-quality sparse correspondences on static regions using SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023), explicitly masking out dynamic objects to ensure geometric validity. We then computed the geometric consistency (via Sampson error) for poses estimated by our pipeline compared to the DroidCalib (Hagemann et al., 2023) baseline.

The results, summarized in Tab. 12, demonstrate the superiority of our approach. Our pipeline reduces the mean reprojection error to 1.09 px (from 1.30 px) and improves fine-grained precision, with 78.36% of correspondences across frame pairs falling within a 1-pixel error threshold, compared to 69.85% for the baseline. Beyond numerical metrics, our pipeline produces visibly more consistent point cloud reconstructions and the visualizations in Appendix G.1 have already demonstrated this.

Table 12: **Evaluation of camera pose annotation reliability.** We report the geometric consistency (reprojection error) on 8,345 sampled pairs, comparing our pipeline against the DroidCalib baseline.

Method	Mean Error (px) ↓	$\% < 1$ Pix ↑	$\% < 3$ Pix ↑	$\% < 5$ Pix ↑
DroidCalib (Baseline)	1.30	69.85%	91.42%	96.02%
Ours	**1.09**	**78.36%**	**93.90%**	**96.66%**

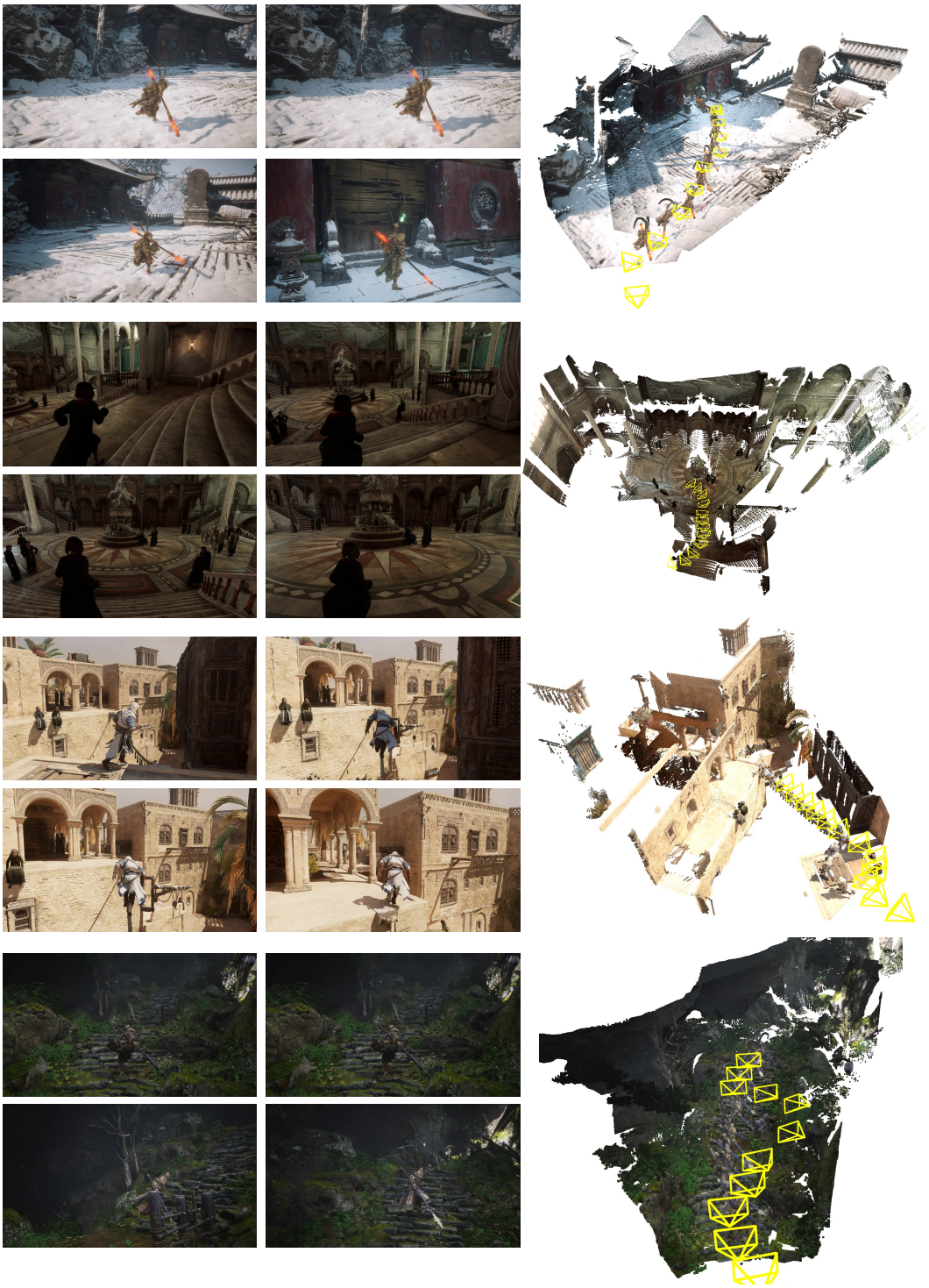


Figure 16: **Visual examples of the *OmniWorld-Game* Benchmark.** Each row displays sample RGB frames from a long sequence alongside the corresponding ground truth 3D point cloud and camera trajectory (indicated by yellow frustums). The samples demonstrate high diversity in scene types and complex camera motions, reflecting the challenging nature of the benchmark.