HOLI-WORLD: BUILDING HOLISTIC WORLD INTELLIGENCE FROM HUMAN-EGO EXPERIENCE

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of HoliWorld.** HoliWorld connects in-the-wild egocentric capture, automated static and dynamic annotation, and vision-language learning. Scene geometry, object semantics, and interaction histories form aligned multimodal supervision for grounding and spatial and temporal reasoning. Representative annotations and results illustrate data coverage, annotation quality, and HoliWorld-VL performance.

ABSTRACT

Human egocentric experience offers a rich foundation for learning how the world is organized and how it changes through interaction. Realizing this potential at scale requires turning growing volumes of recordings into learning resources that connect scene structure, object semantics, and interaction histories across viewpoints and time, while reducing dependence on costly manual annotation. We introduce **HOLIWORLD**, an integrated ecosystem connecting experience capture, world representation, automated data construction, model learning, and independent evaluation. To support learning at scale, **HOLIWORLD-*Foundry*** combines complementary static and dynamic perception to automatically construct world records that connect persistent scene structure and evolving object states with interaction histories and supporting observations. **HOLIWORLD-*Data*** organizes these records from approximately 100 self-collected scenes into grounded learning resources, including approximately 80K QA pairs used to train **HOLIWORLD-*VL-4B***. Independent human annotations of held-out recordings form **HOLIWORLD-*Bench***, supporting evaluation of both annotation quality and learned spatial and temporal understanding. Experiments on ScanNet, ScanNet++, Aria Digital Twin, and HOLIWORLD-*Bench* demonstrate strong performance in static perception and dynamic tracking, while training on HOLIWORLD-*Data* substantially improves HOLIWORLD-*VL-4B*'s spatial and temporal reasoning on HOLIWORLD-*Bench* and EgoDynamic4D. HOLIWORLD provides a practical foundation for expanding learning resources from the continuous flow of human experience and advancing machine understanding of the dynamic world.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Learning how the physical world is organized and how it changes through interaction is a fundamental goal of artificial intelligence and an essential step toward artificial general intelligence (AGI) (Tenenbaum et al., 2011; Lake et al., 2017; LeCun, 2022). Although vision–language learning and predictive video modeling enable machines to recognize objects, describe scenes, interpret activities, and anticipate visual changes (Radford et al., 2021; Li et al., 2023; Liu et al., 2023; Bai et al., 2025; Assran et al., 2025), building a coherent understanding of the world across viewpoints and time remains an open challenge (Yang et al., 2025a; Wang et al., 2026; Lin et al., 2025). Everyday human experience naturally brings these dimensions together: moving through an environment reveals its enduring spatial structure, while interacting with objects reveals how they are used and how their states and relationships change. Therefore, we place human experience at the center of learning world intelligence, asking how it should be captured, organized, and learned from to support an understanding of environments, objects, and interactions as parts of a shared, evolving world.

Egocentric data offers a direct route to learning about the world through the breadth of human life. It records environments as people experience them, bringing together the places they inhabit, the objects they use, and the activities through which their surroundings change (Ragusa et al., 2021; Damen et al., 2022; Tschernezki et al., 2023). Large-scale efforts (Grauman et al., 2022; 2024; Ma et al., 2024; Yang et al., 2025b) have expanded this record across diverse people, places, and activities. However, turning these recordings into detailed learning supervision often requires substantial human input, tying further expansion to additional annotation and curation costs (Darkhalil et al., 2022). The challenge is therefore to scale high-quality supervision with incoming experience while retaining the spatial and temporal information needed to understand environments, objects, and interactions as parts of an evolving world.

Recent advances provide complementary capabilities for automating this process: reconstruction and semantic annotation of predominantly static scenes for spatial supervision (Gao et al., 2026), and egocentric tracking of object identities and 3D locations over time (Zhao et al., 2024; Plizzari et al., 2025). Turning these capabilities into coherent learning resources requires connecting scene geometry and object semantics with interaction histories, then deriving supervision that remains traceable to the underlying observations. To this end, we introduce **HOLIWORLD**, an integrated ecosystem for automatically constructing high-quality learning resources at scale from egocentric recordings that we collect in the wild using our capture device. At its core, **HOLIWORLD-*Foundry*** first reconstructs scene geometry and camera poses, then uses VLM-derived motion and interaction cues to separate static objects from dynamic tracking targets. For static objects, it lifts image detections into 3D and consolidates them with scene point-cloud detections to recover object geometry and semantics. For dynamic objects, it associates identities across interaction segments, tracks object regions in image space, and recovers 3D states and trajectories in shared world coordinates. Built on these multimodal records, **HOLIWORLD-*Data*** organizes approximately 100 scenes and 444 egocentric videos across four levels: raw observations, reconstructed geometry, object and interaction semantics, and spatial intelligence supervision. By further parsing geometric relations, object trajectories, and interaction states, we construct approximately 80K evidence-driven QA pairs covering spatial relations, temporal changes, and interactions, with each answer linked to its supporting observations and annotations. We train **HOLIWORLD-*VL*** on this supervision for visual grounding and spatial reasoning, providing a downstream test of the generated data's learning value. Finally, we establish **HOLIWORLD-*Bench*** through independent human annotation of held-out recordings, providing reference annotations for evaluating both Foundry's world-record accuracy and HOLIWORLD-*VL*'s spatial understanding.

Experiments evaluate HOLIWORLD through both annotation quality and downstream spatial reasoning. For static perception, HOLIWORLD-*Foundry* achieves an AP_{25} of 0.68 on ScanNet (Dai et al., 2017), compared with 0.40 for the strongest baseline in our evaluation. For dynamic tracking, it achieves a $HOTA_{0.1}$ of 54.31% on 20 Aria Digital Twin sequences (Pan et al., 2023), compared with 36.00% for IT3DEgo (Zhao et al., 2024) with ground-truth-guided initialization. The resulting supervision also supports substantial improvements in downstream reasoning: compared with the original Qwen3.5-4B, **HOLIWORLD-*VL-4B*** improves Token F1 from 16.35% to 78.02% on HOLIWORLD-*Bench* and from 18.29% to 66.39% on EgoDynamic4D (Huang et al., 2026).

Our contributions are threefold:

- We introduce **HOLIWORLD**, an integrated ecosystem that provides a practical blueprint for learning world intelligence from human egocentric experience. It connects experience capture, world representation, automated data construction, model learning, and independent evaluation around a shared representation of the evolving world.
- We develop **HOLIWORLD-*Foundry*** to automatically construct world records that preserve persistent scene structure and evolving object states. Complementary static and dynamic perception links geometry, semantics, and interaction histories through persistent identities and supporting observations, enabling scalable construction of grounded supervision.
- We construct **HOLIWORLD-*Data*** to support grounding and spatial and temporal reasoning, and use these resources to train **HOLIWORLD-*VL***. We further establish **HOLIWORLD-*Bench*** through independent human annotation, providing reference data for evaluating both the quality of automatically constructed world records and the capabilities of learned models.

2 RELATED WORK

Egocentric data and annotation. Large-scale egocentric datasets capture diverse activities (Damen et al., 2022; Grauman et al., 2022; 2024). EPIC Fields and Aria Digital Twin add scene geometry (Tschernezki et al., 2023; Pan et al., 2023), while HD-EPIC links actions, object motion, and gaze to 3D digital twins through manual annotation (Perrett et al., 2025). HoliWorld automates the recovery of scene structure and interaction histories from new recordings, reducing the annotation effort needed to expand these learning resources.

Static reconstruction and dynamic tracking. Holi-Spatial constructs static spatial supervision (Gao et al., 2026), and Boxer consolidates multi-view detections into 3D boxes (DeTone et al., 2026). IT3DEgo, EgoSeg3D, and LMK complement these capabilities with object tracking and spatial memory (Zhao et al., 2024; Bhalgat et al., 2024; Plizzari et al., 2025). HoliWorld-Foundry connects persistent scene geometry with evolving object states through shared coordinates, identities, and observation links, supporting spatial and temporal supervision within one representation.

Spatial and temporal understanding. Spatial supervision and geometry-aware representations improve vision–language reasoning (Chen et al., 2024; Ma et al., 2025; Zhu et al., 2025; Zheng et al., 2025). ST-Think and EgoReasoner study egocentric spatial and temporal reasoning (Wu et al., 2026; Zhu et al., 2026), while EgoDynamic4D derives dynamic QA from annotated 4D resources (Huang et al., 2026). We follow EgoDynamic4D's Dynamic QA patterns and address the upstream construction of annotations from raw recordings. HoliWorld connects capture, annotation, supervision generation, and model training, with independent human evaluation of both annotation quality and learned capabilities.

3 METHOD

3.1 HoliWorld Overview

HoliWorld converts egocentric recordings into spatially and temporally aligned supervision (Fig. 2). The Eagle Capture Device records synchronized stereo RGB and IMU measurements; HoliWorld-Foundry reconstructs scenes and annotates static objects and dynamic tracks. HoliWorld-Data organizes these outputs into grounding, captioning, and spatial QA supervision for HoliWorld-VL. Independent human annotations of held-out recordings form HoliWorld-Bench for evaluating Foundry's annotation quality and HoliWorld-VL's spatial understanding.

3.2 HoliWorld Foundry

3.2.1 GEOMETRIC PREPARATION AND SHARED SEMANTIC CUES

As illustrated in Fig. 3, we first recover camera poses and sparse points from stereo RGB and IMU using visual-inertial SLAM, then estimate depth and dense points from RGB keyframes and these poses with Depth Anything 3 (DA3) (Lin et al., 2026). Using RGB observations and this geometry,

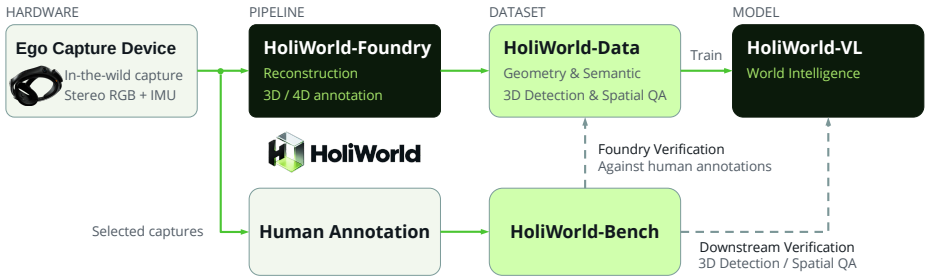


Figure 2: The HOLIWORLD ecosystem connects egocentric capture, Foundry, Data, and VL. Human-annotated Bench evaluates annotation quality and spatial understanding.

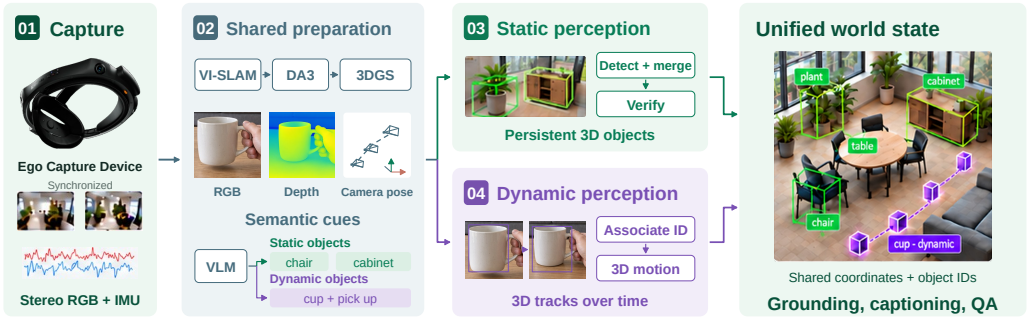


Figure 3: HoliWorld-Foundry overview. Shared preparation reconstructs geometry and routes objects to static or dynamic perception. Static perception (03) detects, merges, and verifies persistent 3D objects; dynamic perception (04) associates identities and recovers 3D tracks. Both outputs are fused into a unified world state.

we optimize a 3D Gaussian scene representation Kerbl et al. (2023) $\mathcal{G}$, render refined depth at the observed poses, and complete dynamic regions with per-frame DA3 depth estimation. We retain $\mathcal{G}$ and the reconstructed scene point cloud $\mathcal{P}$, and organize the resulting observations as

$$\mathcal{O} = \{(I_t, D_t, \mathbf{K}_t, \mathbf{T}_t)\}_{t=1}^{T}, \tag{1}$$

where I_t is an RGB frame, D_t is a depth map, $\mathbf{K}_t$ is the camera intrinsic matrix, and $\mathbf{T}_t$ maps camera coordinates to world coordinates. To route static and dynamic objects to their respective perception pipelines, a vision–language model (VLM) extracts categories, attributes, motion cues, and interaction descriptions from RGB keyframes in $\mathcal{O}$; categories guide static image-based proposals, while motion and interaction cues identify dynamic tracking targets.

3.2.2 STATIC OBJECT PERCEPTION

Dual-path 3D detection. To capture both object details and scene-level geometry, we combine image-based and point-cloud-based 3D detection (Fig. 4A). In the image-based path, we run Boxer (DeTone et al., 2026) separately on the left and right RGB streams to capture complementary object details. Open-vocabulary 2D detections are lifted into metric 3D boxes using the corresponding depth and camera poses, then associated across frames. In parallel, TR3D (Rukhovich et al., 2023) detects objects in the reconstructed scene point cloud $\mathcal{P}$, capturing large-object extents beyond individual views.

Box merging and geometric filtering. Across these complementary sources, multiple proposals may refer to the same object, while others have little geometric support. We therefore use low-cost geometric checks to merge spatially compatible boxes and discard those with insufficient point support in $\mathcal{P}$ before VLM validation (Fig. 4B). For validation, we also preserve one fallback per selected box when available: the rank-1 alternative in its candidate cluster, or, if unavailable, a directly suppressed candidate linked by IoU suppression. Each box retains its own supporting evidence.

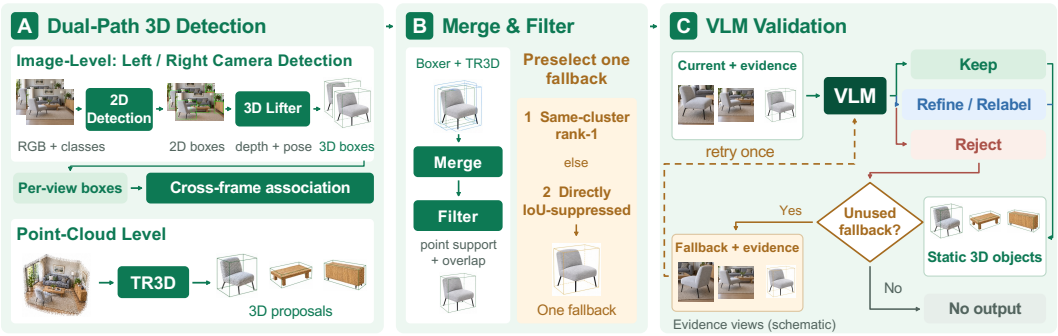


Figure 4: Static object perception. (A) Complementary image and point-cloud paths generate 3D candidates. (B) Geometric merging and filtering retain a current candidate and one preselected fallback. (C) VLM validation uses each candidate's own evidence, with at most one fallback retry. Synthetic illustrations depict inputs and intermediate states.

VLM validation. The remaining boxes may still have inaccurate extents or incorrect category labels. A VLM therefore checks each selected box against source observations and supplementary views rendered from $\mathcal{G}$ (Fig. 4C). Based on this evidence, it can keep the box, refine its geometry, relabel its category, or reject it. If the selected box is rejected, we evaluate its preselected fallback once using the fallback's own evidence. If no unused fallback is available or the retry fails, no box is retained. Accepted boxes form $\mathcal{B}^{\text{sta}}$ with links to their supporting observations.

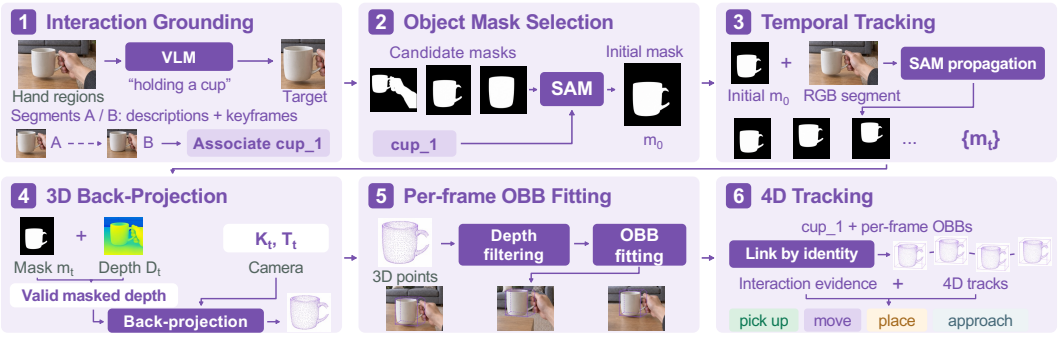


Figure 5: Dynamic object perception. Cross-segment identity association guides mask selection and within-segment propagation. Valid masked depth yields filtered 3D points and per-frame 3D oriented bounding boxes (OBBs). Linking boxes by identity produces 4D tracks; tracks and interaction evidence jointly support motion cues. Synthetic illustrations depict inputs and intermediate states.

3.2.3 Dynamic Object Perception

Interaction grounding and identity association. Manipulated objects may disappear from view and reappear in later interaction segments. To recognize the same object across these interruptions, we use hand regions and VLM interaction descriptions to identify the target in each segment, then compare descriptions and keyframe evidence across segments. This *temporal label continuity* establishes consistent object identities that guide target mask selection (Fig. 5).

Identity-guided mask tracking. The associated identities specify which object to follow, while frame-level masks are needed to localize it over time. Guided by these identities, the Segment Anything Model (SAM) (Kirillov et al., 2023; Ravi et al., 2025) selects and initializes the target mask, then propagates it within each interaction segment to obtain $\{m_t\}$. We assume at most one moving object at each time, with m_t denoting its foreground pixels.

3D recovery and motion cues. Image-space motion also reflects camera movement. To recover object states in world coordinates, we back-project the pixels of each valid tracked mask using the DA3-completed dynamic depths in D_t and the camera pose:

$$\mathcal{X}_t = \left\{ \mathbf{T}_t \left(D_t(\mathbf{u}) \mathbf{K}_t^{-1} \bar{\mathbf{u}} \right) \mid \mathbf{u} \in m_t,\ D_t(\mathbf{u}) \text{ is valid} \right\}, \tag{2}$$

where $\bar{\mathbf{u}}$ is the homogeneous image coordinate, $\mathbf{T}_t(\cdot)$ applies the camera-to-world rigid transformation, and $\mathcal{X}_t$ contains the object-supported world points. After filtering unsupported depth samples, we fit an OBB to each point set and link boxes with the same identity across time to form a dynamic 4D track.

3.2.4 Unified World Representation

Object records connect instance identities and categories to image evidence, 3D geometry, and interaction cues. We express the Foundry output at time t as

$$\mathcal{S}(t) = \big(\mathcal{O}_t, \mathcal{U}(t), \mathcal{G}, \mathcal{P}, \mathcal{B}^{\text{sta}}, \mathcal{B}^{\text{dyn}}(t), \mathcal{Z}(t)\big), \tag{3}$$

where $\mathcal{O}_t = (I_t, D_t, \mathbf{K}_t, \mathbf{T}_t)$ is the input observation, $\mathcal{U}(t)$ contains the detected 2D boxes and tracked masks, and $\mathcal{Z}(t)$ records the associated labels, identities, confidence scores, interaction semantics, and evidence links. Static instances persist across observations; dynamic instances retain time-indexed states. This representation lets supervision refer to specific objects, times, and supporting observations.

3.3 HoliWorld-Data

HoliWorld-Data organizes the world records constructed by Foundry into grounding, captioning, and spatial, temporal, and interaction QA supervision. The available evidence determines the scope of these tasks: static objects and their geometry support spatial understanding; dynamic tracks with consistent object identities support motion and temporal understanding; and tracks paired with interaction semantics support descriptions of object manipulation and state changes. Each example retains links to source observations and annotations, allowing its target output to be traced to specific objects, times, and supporting evidence.

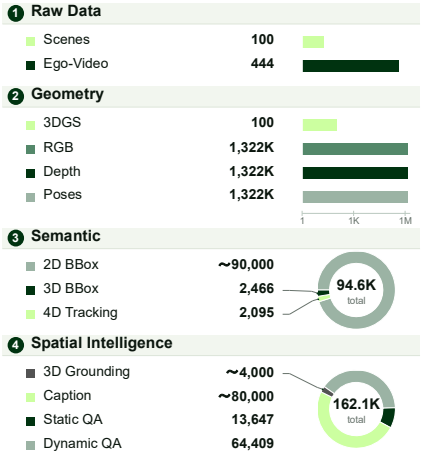


Figure 6: HoliWorld-Data composition and its spatial-temporal supervision.

QA pairs construction. We construct QA pairs from the spatial, temporal, and interaction annotations in Foundry's world records, following the QA patterns of EgoDynamic4D (Huang et al., 2026). Grounding tasks associate referring expressions with the corresponding 2D observations, static 3D boxes, or dynamic tracks. Spatial QA uses object geometry and camera poses to derive direction, distance, containment, support, and viewpoint-dependent relations. Temporal QA compares the states of the same object over time to describe motion, event ordering, and duration. Interaction QA additionally uses interaction semantics and source observations to describe manipulated objects and state changes. Scene captions and inventory questions aggregate object records across frames to describe the observed environment. Dynamic supervision follows Foundry's current assumption of at most one moving object at each time.

Dataset composition. HoliWorld-Data contains roughly 100 scenes and 444 egocentric videos, with approximately 100 3D Gaussian splatting (3DGS) reconstructions, 1,322K RGB frames with corresponding depth maps and camera poses, 90K 2D boxes, 2,466 static 3D boxes, and 2,095 dynamic tracks (Fig. 6). These records support about 4K 3D grounding annotations and 80K captions, together with 13,647 static QA pairs and 64,409 dynamic QA pairs, totaling about 80K QA pairs.

4 EXPERIMENTS

We evaluate the pipeline's ability to generate object-centric 3D annotations for both static scenes and moving objects. Sections 4.1 and 4.2 present the results for static object perception and dynamic object tracking. Section 4.4 examines the contributions of key pipeline components through controlled ablations. Detailed experimental settings are provided in Appendix Sections B.1 and B.2.

4.1 Static 3D Object Perception

As shown in Table 1, HOLIWORLD-FOUNDRY achieves the highest AP_{25} and AP_{50} on all three datasets. On ScanNet (Dai et al., 2017) and ScanNet++ (Yeshwanth et al., 2023), AP_{50} reaches 0.55 and 0.13, respectively, compared with 0.12 and 0.05 for the strongest baselines, indicating more accurate detection under stricter geometric matching. The method also leads on the more complex real-world scenes in HoliWorld-Bench. Fig. 7 shows more complete object coverage and better-aligned box extents across representative scenes, consistent with these quantitative gains.



Figure 7: Qualitative static 3D object perception results across representative scenes.

Table 1: Static 3D perception on ScanNet, ScanNet++, and HoliWorld-Bench ($[0, 1]$ scale). AP uses 3D IoU thresholds of 0.25/0.50; mIoU uses pairs matched at IoU 0.25.

	ScanNet			ScanNet++			HoliWorld-Bench		
Method	AP_{25} ↑	AP_{50} ↑	mIoU↑	AP_{25} ↑	AP_{50} ↑	mIoU↑	AP_{25} ↑	AP_{50} ↑	mIoU↑
Model-based methods									
SpatialLM	0.18	0.09	0.52	0.03	0.01	0.50	0.04	0.00	0.23
LLaVA-3D	0.02	0.00	0.34	0.00	0.00	0.34	0.00	0.00	0.00
SceneScript	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Map-Det3D	0.40	0.12	0.48	0.18	0.05	0.48	0.12	0.01	0.37
3DGS-based methods									
M3-Spatial	0.00	0.00	0.37	0.00	0.00	0.60	0.06	0.00	0.28
LangSplat	0.02	0.00	0.45	0.01	0.00	0.33	0.01	0.00	0.10
Spatial pipeline frameworks									
Boxer	0.32	0.09	0.47	0.23	0.05	0.46	0.16	0.07	0.52
Holi-Spatial	0.33	0.10	0.47	0.17	0.05	0.45	0.07	0.01	0.37
Ours	0.68	0.55	0.72	0.30	0.13	0.56	**0.24**	**0.12**	**0.53**

4.2 Dynamic 3D Object Perception and Tracking

In Table 2, HOLIWORLD-FOUNDRY attains HOTA scores of 54.31% on ADT and 43.14% on HoliWorld-Bench, compared with 36.00% and 15.13% for the strongest baselines, respectively. Higher AssA and PCL indicate more consistent identities and broader trajectory coverage across both datasets. Box metrics provide a complementary comparison of spatial extents on matched objects. Fig. 8 illustrates trajectories that more closely follow the reference paths across the shown interaction stages.



Figure 8: Dynamic tracking in three scenes. Solid/dashed curves show predicted/GT trajectories; panels show initial, middle, and final observations. Red crosses mark unmatched observations.

Table 2: Dynamic tracking on ADT20 and HoliWorld-Bench (%) at a 0.1 m trajectory tolerance. **GT first-visible mask initialization*; †*GT category queries*. See Appendix B.2 for input adaptations and evaluation protocols.

	ADT20						HoliWorld-Bench					
	Bounding box			Trajectory			Bounding box			Trajectory		
Method	IoU↑	GTC↑	PP↑	HOTA↑	AssA↑	PCL↑	IoU↑	GTC↑	PP↑	HOTA↑	AssA↑	PCL↑
Instance segmentation and tracking pipeline												
EgoSeg3D†	–	–	–	21.89	37.11	36.47	–	–	–	5.76	21.78	36.84
IT3DEgo*	–	–	–	36.00	53.60	34.64	–	–	–	2.44	4.36	7.89
SAM+Whareformer*	–	–	–	30.76	40.99	31.26	–	–	–	14.63	20.97	36.84
3D object perception and tracking pipelines												
SAM+LMK*	7.96	21.27	14.53	26.10	30.76	21.45	20.28	50.84	21.30	15.13	14.35	23.68
Ours	**30.88**	**43.03**	**56.25**	**54.31**	**71.94**	**47.95**	**41.79**	**53.18**	**76.49**	**43.14**	**56.54**	**57.89**

4.3 Downstream Spatial and Temporal Understanding

Fine-tuning Qwen3.5-4B (Qwen Team, 2026a) on HoliWorld-Data raises Token F1 from 16.35% to 78.02% on HoliWorld-Bench and from 18.29% to 66.39% on EgoDynamic4D (Table 3), supporting the value of the generated supervision for spatial and temporal reasoning. The resulting HoliWorld-VL-4B also surpasses zero-shot Qwen3.8-27B (Qwen Team, 2026b) on all four metrics across both benchmarks. The gains in yes/no, multiple-choice, and numerical accuracy complement the Token F1 improvements, indicating that the generated supervision benefits multiple question formats across the two benchmarks.

Table 3: Spatial reasoning results (%). F1 denotes Token F1; Y/N, MCQ, and Num. report yes/no, multiple-choice, and numerical accuracy.

Model	F1	Y/N	MCQ	Num.
Holi-World benchmark				
Qwen3.5-4B	16.35	34.86	62.00	10.80
Qwen3.8-27B	43.67	82.85	75.00	66.13
HoliWorld-VL-4B	**78.02**	**93.52**	**94.00**	**81.84**
EgoDynamic4D				
Qwen3.5-4B	18.29	59.47	79.16	0.02
Qwen3.8-27B	42.93	61.07	57.54	0.16
HoliWorld-VL-4B	**66.39**	**99.67**	**94.92**	**16.33**

4.4 Ablation Studies

Static components. Fig. 9(a–b) illustrates complementary geometric coverage: image-based lifting method captures more small objects, point-cloud detection recovers larger extents, and stereo fusion combines observations from both views. Caption guidance recovers missed entities (c). VLM verification shows modest changes in geometric metrics, while the qualitative examples illustrate

corrections to object labels and unsuitable candidates (d), suggesting a complementary role in semantic annotation. Detailed AP_{25} and mIoU comparisons are reported in Table 5 (Appendix B.3).

Dynamic components. Fig. 9(e–f) highlights two complementary requirements for reliable trajectories. Cross-interaction label continuity reduces identity fragmentation across repeated observations, while SAM association supplies temporally consistent masks for 3D lifting, reducing the trajectory deviations seen with per-frame detection. Together, they preserve both object correspondence and spatial continuity. Corresponding HOTA and PCL gains on both datasets are reported in Table 5 (Appendix B.3).

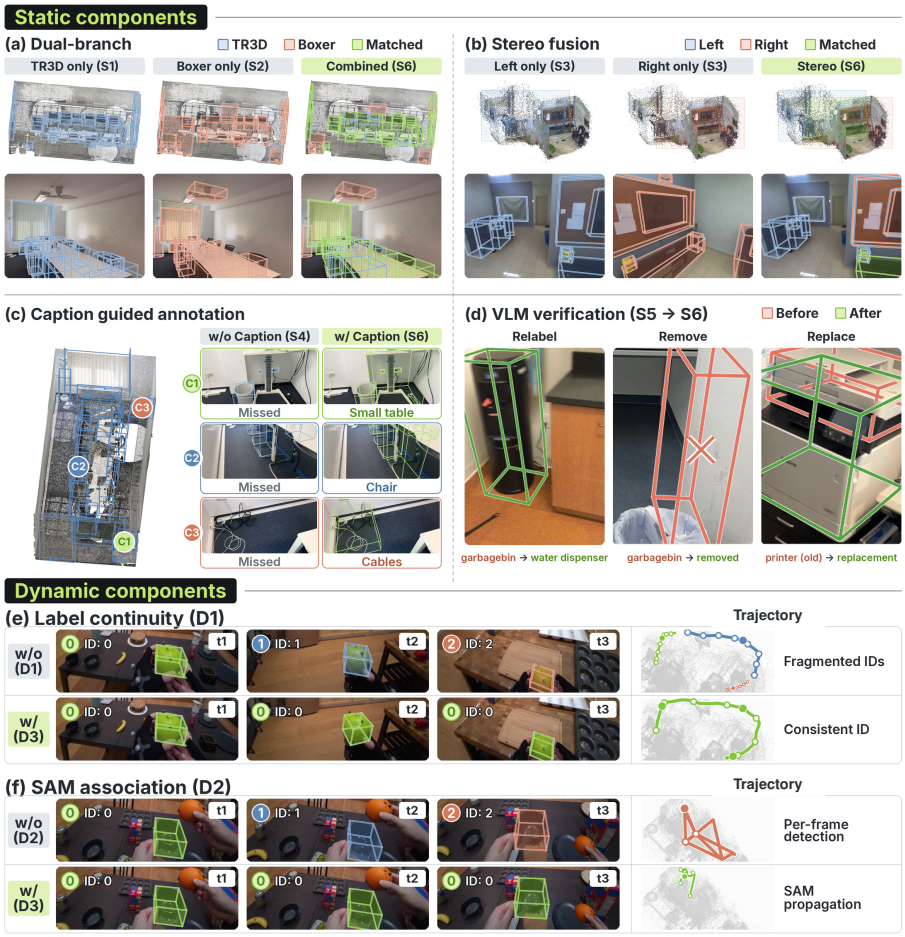


Figure 9: Qualitative ablation results for the static and dynamic components.

5 Conclusion

In this paper, we presented HoliWorld, an integrated ecosystem that places human egocentric experience at the center of learning world intelligence. Through complementary static and dynamic perception, HoliWorld-*Foundry* automatically constructs world records linking scene geometry, object semantics, and interaction histories to supporting observations. These records form HoliWorld-*Data* and support HoliWorld-*VL* training, while independently annotated HoliWorld-*Bench* provides reference data for evaluating annotation quality and learned capabilities. Experiments demonstrate improved static perception and dynamic tracking, alongside gains in spatial and temporal reasoning on HoliWorld-*Bench* and EgoDynamic4D. HoliWorld provides a practical foundation for expanding these grounded learning resources with new recordings, enabling machines to learn how the world is organized and how it changes through interaction.

AI USE STATEMENT

We used generative AI tools solely to polish the language and improve the clarity, conciseness, and readability of the manuscript. The authors reviewed and verified all AI-assisted text and take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, Section 3 details the HoliWorld pipeline, covering geometric reconstruction, static object perception, dynamic tracking, and supervision construction, alongside dataset composition and QA construction. Appendix B further specifies evaluation datasets, metric definitions, baseline adaptations, matching protocols, and ablation configurations.

ETHICS STATEMENT

HoliWorld is designed for research on scene geometry and spatial-temporal understanding. Responsible data collection and release therefore require appropriate consent, review and redaction of identifying content, and compliance with dataset and model licenses. The pipeline should not be used for non-consensual surveillance or high-stakes decisions without additional safeguards and validation.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. SceneScript: Reconstructing scenes with an autoregressive structured language model. In *Computer Vision – ECCV 2024*, pp. 247–263, 2025. doi: 10.1007/978-3-031-73030-6_14. URL https://link.springer.com/chapter/10.1007/978-3-031-73030-6_14.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yash Bhalgat, Vadim Tschernezki, Iro Laina, Joao F. Henriques, Andrea Vedaldi, and Andrew Zisserman. 3D-aware instance segmentation and tracking in egocentric videos. In *Asian Conference on Computer Vision*, 2024.

Jacob Chalk, Saptarshi Sinha, Dima Damen, Yannis Kalantidis, and Diane Larlus. Whareformer: Learning to track what is where in long egocentric videos. In *European Conference on Computer Vision (ECCV)*, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cheng_Tracking_Anything_with_Decoupled_Video_Segmentation_ICCV_2023_paper.html.

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 (CVPR)*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022. doi: 10.1007/s11263-021-01531-2.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2209.13064.

Daniel DeTone, Tianwei Shen, Fan Zhang, Lingni Ma, Julian Straub, Richard Newcombe, and Jakob Engel. Boxer: Robust lifting of open-world 2D bounding boxes to 3D. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://github.com/facebookresearch/boxer.

Yuanyuan Gao, Hao Li, Yifei Liu, Xinhao Ji, Yuning Gong, Yuanjun Liao, Fangfu Liu, Manyuan Zhang, Yuchen Yang, Dan Xu, Xue Yang, Huaxi Huang, Hongjie Zhang, Ziwei Liu, Xiao Sun, Dingwen Zhang, and Zhihang Zhong. Holi-Spatial: Evolving video streams into holistic 3D spatial intelligence. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://visionary-laboratory.github.io/holi-spatial/.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, 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*, 2024. URL https://arxiv.org/abs/2311.18259.

Junsheng Huang, Shengyu Hao, Bo-Cheng Hu, Hongwei Wang, and Gaoang Wang. Understanding dynamic scenes in egocentric 4D point clouds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5031–5039, 2026. doi: 10.1609/aaai.v40i7.37416. URL https://ojs.aaai.org/index.php/AAAI/article/view/37416.

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

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Yann LeCun. A path towards autonomous machine intelligence, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf. Position paper, version 0.9.2.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, Wenbo Hu, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Video-Bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pp. 38–55, 2025. doi: 10.1007/978-3-031-72970-6_3. URL https://link.springer.com/chapter/10.1007/978-3-031-72970-6_3.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. doi: 10.1007/s11263-020-01375-2.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2406.09905.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. SpatialLLM: A compound 3D-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17249–17260, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_SpatialLLM_A_Compound_3D-Informed_Design_towards_Spatially-Intelligent_Large_Multimodal_Models_CVPR_2025_paper.html.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. SpatialLM: Training large language models for structured indoor modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/40223e204a9e18dd8b01d7d11ea97939-Abstract-Conference.html.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria Digital Twin: A new benchmark dataset for egocentric 3D machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20133–20143, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Pan_Aria_Digital_Twin_A_New_Benchmark_Dataset_for_Egocentric_3D_ICCV_2023_paper.html.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23901–23913, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Perrett_HD-EPIC_A_Highly-Detailed_Egocentric_Video_Dataset_CVPR_2025_paper.html.

Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *International Conference on 3D Vision (3DV)*, 2025.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The MECCANO dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1569–1578, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Ragusa_The_MECCANO_Dataset_Understanding_Human-Object_Interactions_From_Egocentric_Videos_in_WACV_2021_paper.html.

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. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Tr3d: Towards real-time indoor 3d object detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 281–285. IEEE, 2023.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. doi: 10.1126/science.1192788.

Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D geometry and video understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.08731.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Fei-Fei Li, and Manling Li. Spatial mental modeling from limited views. In *International Conference on Learning Representations*, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0457. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/142b166c8b18b43774c4f0e0c1e9948e-Abstract-Conference.html.

Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. ST-Think: How multimodal large language models reason about 4D worlds from ego-centric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5174–5183, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Wu_ST-Think_How_Multimodal_Large_Language_Models_Reason_About_4D_Worlds_WACV_2026_paper.html.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025a.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_EgoLife_Towards_Egocentric_Life_Assistant_CVPR_2025_paper.html.

Yung-Hsu Yang, Luigi Piccinelli, Samuel Rota Bulò, Sunghwan Hong, Denis Rozumny, Johannes Schönberger, Zuria Bauer, Hermann Blum, Peter Kontschieder, and Marc Pollefeys. MapDet3D: Metric feed-forward 3D reconstruction prior for multi-view 3D object detection from streaming inputs. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://royyang0714.github.io/Map-Det3D/.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. URL https://scannetpp.mlsg.cit.tum.de/scannetpp/.

Yunhan Zhao, Haoyu Ma, Shu Kong, and Charless Fowlkes. Instance tracking in 3D scenes from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zheng_Video-3D_LLM_Learning_Position-Aware_Video_Representation_for_3D_Scene_Understanding_CVPR_2025_paper.html.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3D capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_LLaVA-3D_A_Simple_yet_Effective_Pathway_to_Empowering_LMMs_with_ICCV_2025_paper.html.

Fangrui Zhu, Yunfeng Xi, Jianmo Ni, Mu Cai, Boqing Gong, Long Zhao, Chen Qu, Ian Miao, Yi Li, Cheng Zhong, Huaizu Jiang, and Shwetak Patel. EgoReasoner: Learning egocentric 4D

reasoning via task-adaptive structured thinking. *arXiv preprint arXiv:2603.06561*, 2026. URL https://arxiv.org/abs/2603.06561.

Xueyan Zou, Yuchen Song, Ri-Zhao Qiu, Xuanbin Peng, Jianglong Ye, Sifei Liu, and Xiaolong Wang. 3D-spatial multimodal memory. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/639f9db06bf70c4ff44cf22a7f92cb08-Abstract-Conference.html.

A Related Work

A.1 Egocentric Datasets and Annotations

Large-scale egocentric datasets cover diverse human activities and environments (Damen et al., 2022; Grauman et al., 2022; 2024). EPIC Fields and Aria Digital Twin add camera and scene geometry (Tschernezki et al., 2023; Pan et al., 2023), while HD-EPIC (Perrett et al., 2025) connects actions, object movements, and gaze to 3D digital twins through detailed manual annotation. However, extending such detailed supervision to new recordings entails further annotation effort (Darkhalil et al., 2022; Perrett et al., 2025). HoliWorld addresses this scaling bottleneck by automatically recovering scene structure, object states, and interaction annotations from newly captured recordings and converting them into learning supervision, reducing dependence on manual labeling as new experience is collected.

A.2 3D Scene Reconstruction and Object Tracking

Holi-Spatial (Gao et al., 2026) automates static scene reconstruction, semantic annotation, and spatial QA generation at scale, while Boxer (DeTone et al., 2026) consolidates multi-view detections into static 3D boxes. However, static scene consolidation does not retain the successive object states needed to describe motion and interactions. Egocentric tracking provides complementary capabilities: IT3DEgo and EgoSeg3D (Zhao et al., 2024; Bhalgat et al., 2024) maintain object identities and locations over time, and LMK, introduced for the Out of Sight, Not Out of Mind (OSNOM) task (Plizzari et al., 2025), maintains spatial memory beyond the current view. HoliWorld-*Foundry* connects these static and dynamic perception objectives within an automated data construction pipeline. Shared world coordinates and persistent identities link scene geometry, evolving object states, and interaction evidence, enabling spatial and temporal supervision to refer to the same objects and supporting observations.

A.3 Spatial and Temporal Reasoning in Vision–Language Models

Dedicated spatial supervision (Chen et al., 2024; Ma et al., 2025) and geometry-aware visual representations (Zhu et al., 2025; Zheng et al., 2025; Wu et al., 2025) improve spatial reasoning, while video benchmarks (Yang et al., 2025a; Lin et al., 2025) test these capabilities across observations. ST-Think and EgoReasoner (Wu et al., 2026; Zhu et al., 2026) further study reasoning over egocentric spatial and temporal information. EgoDynamic4D (Huang et al., 2026) constructs dynamic QA from existing annotated 4D resources and introduces a model integrating static and dynamic scene information. We follow its QA patterns, while addressing the upstream challenge of automatically constructing the underlying annotations from new recordings. HoliWorld connects experience capture, automated annotation, supervision generation, and model training, allowing new recordings to expand the learning resources. Independent human-annotated evaluation then assesses both the accuracy of the recovered records and the capabilities learned from them.

B Experimental Details

B.1 Static Perception Evaluation Details

Data and metrics. We evaluate on 20 sequences from each of ScanNet and ScanNet++, using their official GT 3D annotations, and 20 self-collected in-house sequences with manually annotated GT. AP_{25} and AP_{50} use 3D IoU thresholds of 0.25 and 0.50, respectively. Matched-box mIoU is averaged over box pairs matched at IoU 0.25. All static metrics use a $[0, 1]$ scale; dashes indicate unavailable results.

Baselines. The comparison covers model-based methods, including the generalist scene models SpatialLM (Mao et al., 2025), LLaVA-3D (Zhu et al., 2025), and SceneScript (Avetisyan et al., 2025), and the specialist detector Map-Det3D (Yang et al., 2026); 3DGS-based methods M3-Spatial (Zou et al., 2025) and LangSplat (Qin et al., 2024), which encode semantics in Gaussian representations; and spatial pipelines Boxer (DeTone et al., 2026) and Holi-Spatial (Gao et al., 2026), which integrate multiple perception and geometric processing stages.

Additional analysis. At IoU 0.50, HOLIWORLD-FOUNDRY reaches AP_{50} of 0.55 on ScanNet and 0.13 on ScanNet++, compared with the strongest baseline scores of 0.12 and 0.05 (Table 1). These gains persist under stricter geometric matching. On the in-house data, the improvement over Boxer is more pronounced in AP than in matched-box mIoU, suggesting that improved detection contributes beyond the extent accuracy of successfully matched boxes. Because mIoU is conditional on matching, it should be read jointly with AP when assessing overall perception quality.

B.2 Full Dynamic Tracking Results

Table 4 and Figure 11 report the full quantitative and qualitative comparisons.

Metrics. $HOTA_{0.1}$, $DetA_{0.1}$, and $AssA_{0.1}$ follow Luiten et al. (2021), using a 0.1 m world-coordinate center-distance tolerance. They measure joint detection and association quality, detection accuracy, and identity association consistency, respectively. $PCL_{0.1}$ measures correctly localized GT instance-frames under a fixed, sequence-global one-to-one GT–prediction assignment that maximizes matches within 0.1 m; unassigned trajectories, missing predictions, and localization failures count as incorrect. Recall $R_{0.1}$, precision $P_{0.1}$, and TP-L2 use HOTA's frame-level matches to measure GT coverage, prediction reliability, and mean center error among successful matches. Box geometry is assessed separately using IoU, GT coverage (GTC), and prediction precision (PP). All metrics are percentages except TP-L2, which is in meters.

Baselines and inputs. EgoSeg3D† (Bhalgat et al., 2024) receives sequence-level GT category queries. Its GroundingDINO–SAM-v1–DEVA frontend (Liu et al., 2025; Kirillov et al., 2023; Cheng et al., 2023) generates and propagates masks before 3D lifting and native association. For IT3DEgo* (Zhao et al., 2024), we provide a predefined target set and convert each target's first-visible GT mask to a bounding box for its native SVOE enrollment interface; subsequent mask/proposal inference and memory updates use the original pipeline.

SAM+Whareformer* (Chalk et al., 2026) and SAM+LMK* (Plizzari et al., 2025) share a SAM2 propagation frontend (Ravi et al., 2025), initialized with each target's GT mask at its first visible frame in the video, irrespective of when it starts moving. Propagation continues through the remaining video to supply mask observations for native association. An adapter derives Whareformer's world centers from masks, depth, and camera poses; LMK retains its native 3D center/scale outputs. Ours uses no annotation-derived test-time inputs. GT initialization defines baseline inputs only, not evaluation correspondences.

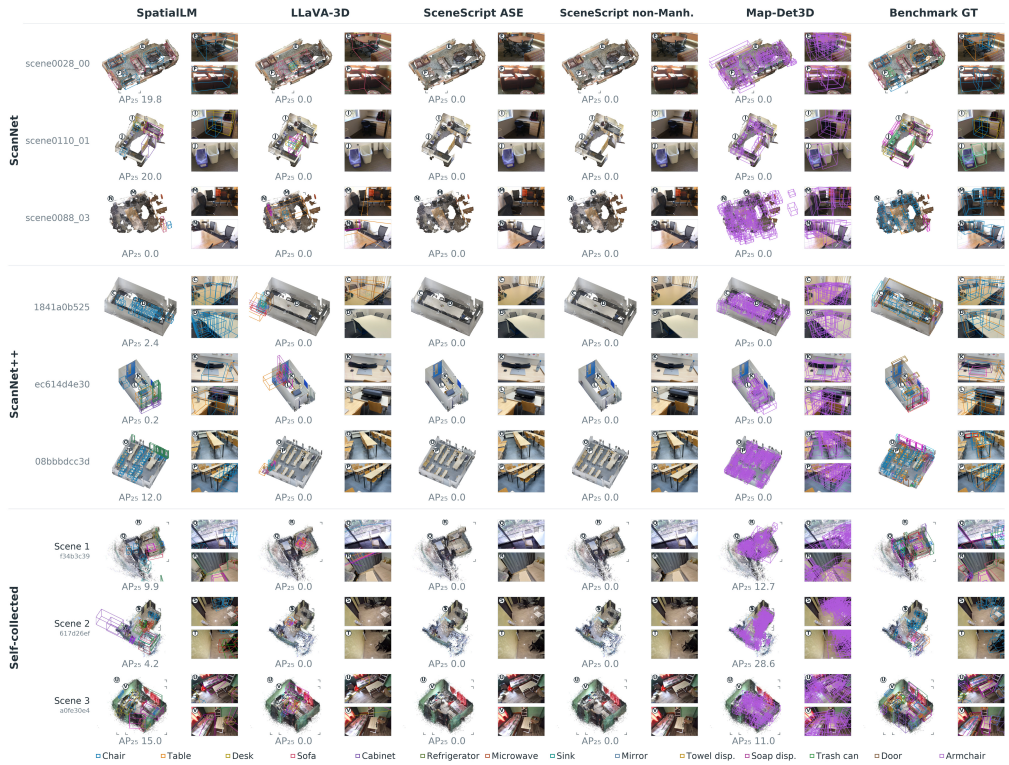
Evaluation and binding rules. We evaluate 20 sequences each from ADT and HoliWorld-Bench over the full RGB timeline, including out-of-view periods, wherever a valid GT pose is available within 20 ms. Missing predictions remain evaluation failures, and sequence scores are averaged equally. We group pipelines by whether they output 3D boxes. Output-dependent GT–prediction binding uses nearest world-coordinate centers; for pipelines with 3D boxes, candidate tracks must additionally have maximum 3D IoU > 0.5. IoU, GTC, and PP are computed on the resulting bound-object subset. This binding does not filter trajectory evaluation: all methods use the common distance-based scoring described above, with native predicted identities and PCL's own sequence-global assignment.

B.3 Additional Ablation Results

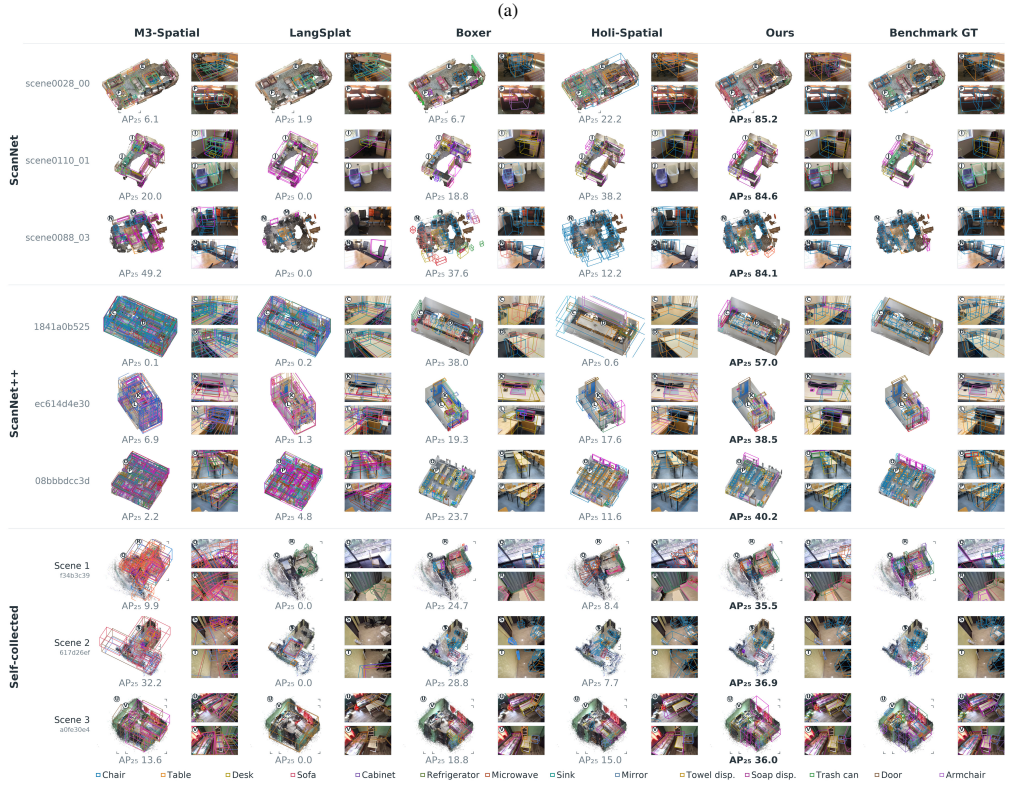
Table 5 reports component-wise results; Figure 12 compares the two static perception branches.

B.4 Holi-World Static and Dynamic QA

30 QA types. We design 30 fine-grained QA types, broadly divided into static and dynamic reasoning. The static category contains 11 types that evaluate camera motion and object-level spatial relationships, including distance, depth, height, and relative position under single- and multi-view settings. The dynamic category comprises seven trajectory-based types for motion status, timing, direction, sequence, and path characteristi cs, together with 12 caption-based types covering action recognition, object interaction, temporal grounding, interaction phases, event ordering, state transitions, and contrastive hard negatives. As illustrated in Fig. 13, these questions combine visual observations with 3D bounding boxes, object trajectories, and temporally aligned captions, while their answers span categorical labels, numerical values, multiple-choice options, time intervals, and free-form descriptions.



(a)



(b)

Figure 10: Qualitative comparison of static 3D object perception across methods and scenes. Panels (a) and (b) show complementary subsets of the evaluated methods.

Table 4: Full dynamic tracking results on ADT20 and HoliWorld-Bench (%; TP-L2 in meters). **GT first-visible mask initialization* (converted to boxes for IT3DEgo); †*GT category queries*. Ours uses no annotation-derived test-time inputs. Box metrics use bound objects; dashes denote unavailable results.

	Bounding box			Trajectory						
Method	IoU↑	GTC↑	PP↑	$HOTA_{0.1}$ ↑	$DetA_{0.1}$ ↑	$AssA_{0.1}$ ↑	$PCL_{0.1}$ ↑	$R_{0.1}$ ↑	$P_{0.1}$ ↑	TP-L2↓
ADT20										
2D instance tracking pipelines										
EgoSeg3D†	–	–	–	21.89	13.78	37.11	36.47	46.96	16.67	0.0432
IT3DEgo*	–	–	–	36.00	24.63	53.60	34.64	35.69	40.93	0.0490
SAM+Whareformer*	–	–	–	30.76	23.60	40.99	31.26	33.35	41.33	0.0497
3D object perception and tracking pipelines										
SAM+LMK*	7.96	21.27	14.53	26.10	22.84	30.76	21.45	27.87	52.17	0.0485
Ours	**30.88**	**43.03**	**56.25**	**54.31**	**42.13**	**71.94**	**47.95**	**49.56**	**76.68**	**0.0333**
HoliWorld-Bench (in-house)										
2D instance tracking pipelines										
EgoSeg3D†	–	–	–	5.76	1.52	21.78	36.84	42.11	1.55	**0.0094**
IT3DEgo*	–	–	–	2.44	1.36	4.36	7.89	7.89	1.62	0.0479
SAM+Whareformer*	–	–	–	14.63	10.20	20.97	36.84	39.47	12.10	0.0113
3D object perception and tracking pipelines										
SAM+LMK*	20.28	50.84	21.30	15.13	15.96	14.35	23.68	39.47	21.13	0.0113
Ours	**41.79**	**53.18**	**76.49**	**43.14**	**32.91**	**56.54**	**57.89**	**68.42**	**38.81**	0.0280



Figure 11: Qualitative dynamic tracking comparison across all evaluated methods.

Table 5: Component ablations. Static AP_{25}/mIoU: $[0, 1]$; dynamic HOTA/PCL: % at 0.1 m. Checkmarks/crosses: enabled/disabled; dashes: unavailable or inapplicable.

	Static components					Dynamic components		Static data				Dynamic data			
								ScanNet		HoliWorld-Bench		ADT20		HoliWorld-Bench	
ID	Boxer	TR3D	StereoGeo	Caption	VLM-Ver.	Label-Cont.	SAM-Assoc.	AP_{25} ↑	mIoU↑	AP_{25} ↑	mIoU↑	HOTA↑	PCL↑	HOTA↑	PCL↑
S1	×	✓	✓	✓	✓	–	–	0.58	0.80	0.10	0.47	–	–	–	–
S2	✓	×	✓	✓	✓	–	–	0.32	0.30	0.20	0.52	–	–	–	–
S3	✓	✓	×	✓	✓	–	–	–	–	0.20	0.53	–	–	–	–
S4	✓	✓	✓	×	✓	–	–	0.67	0.70	0.21	0.52	–	–	–	–
S5	✓	✓	✓	✓	×	–	–	0.68	0.72	0.24	0.52	–	–	–	–
S6	✓	✓	✓	✓	✓	–	–	0.68	0.72	0.24	0.53	–	–	–	–
D1	–	–	–	–	–	×	✓	–	–	–	–	44.09	42.62	32.86	45.30
D2	–	–	–	–	–	✓	×	–	–	–	–	38.38	34.08	38.89	51.04
D3	–	–	–	–	–	✓	✓	–	–	–	–	54.31	47.95	43.14	57.89

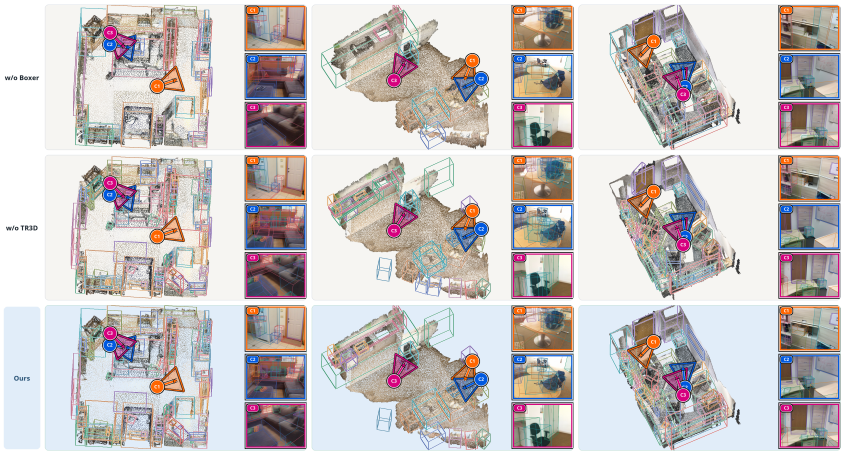


Figure 12: Qualitative ablation of dual-branch static object perception across three scenes. Rows show results without 3D lifting (w/o Boxer), without point-cloud detection (w/o TR3D), and with both branches (Ours). C1–C3 identify camera viewpoints and their corresponding image projections.

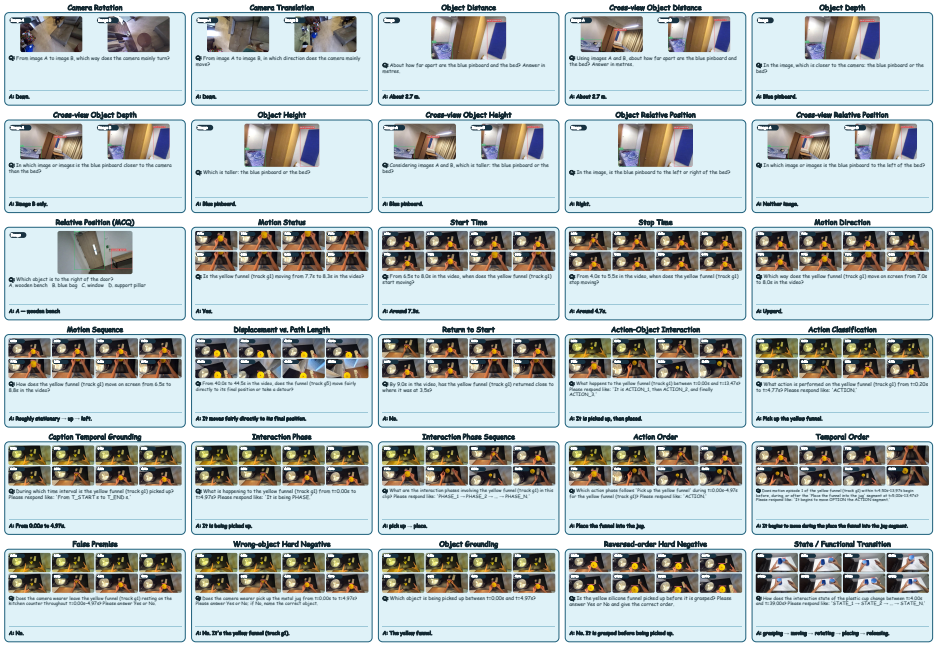


Figure 13: Examples of the 30 QA types in Holi-World, covering static mapping, dynamic trajectory, and dynamic caption reasoning.