MULTIVIEW-EB: BENCHMARKING EGO–EXO EMBODIED UNDERSTANDING OF MLLMS IN REAL-WORLD SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents rely on egocentric observations for perception and decision-making, yet limited fields of view and frequent occlusions can conceal task-relevant objects, state changes, and spatial relations. Synchronized exocentric views offer complementary evidence, but existing benchmarks focus on single-view reasoning or static cross-view perception. Whether multimodal large language models can integrate both views over time to improve egocentric decisions remains unclear. We introduce MultiView-EB, a benchmark for cross-view spatiotemporal understanding in real-world embodied activities. Built from synchronized recordings of human activities, it contains 2,497 samples across nine activity domains, obtained through semi-automated annotation and human verification. Eight tasks assess cross-view fine-grained perception, task planning, and trajectory prediction through open-ended, multiple-choice, and spatial prediction outputs. We evaluate thirteen general-purpose MLLMs and eight embodied-brain models and analyze performance differences across egocentric and combined-view inputs. On the planning tasks, where input conditions can be matched, adding an exocentric view yields an average gain of only 0.5–2.1 points over egocentric input and degrades accuracy for four of the eight models tested. Additional views thus do not automatically improve embodied reasoning: current models struggle to establish cross-view correspondences, align temporal dynamics, and selectively use complementary evidence for planning. Finally, we introduce MV-EmData-38K, on which fine-tuning improves performance on all eight tasks. MultiView-EB provides a systematic testbed for embodied models capable of reasoning across views.

1 INTRODUCTION

Multimodal large language models (MLLMs) are increasingly used as embodied brains that interpret observations and guide high-level decisions (OpenAI, 2026b; Google, 2025; Anthropic, 2026; Qwen Team, 2026a; Wang et al., 2025; An et al., 2026; Tan et al., 2026; Li et al., 2026; Wang et al., 2026; Liang et al., 2026; Agarwal et al., 2026; Lin et al., 2025; Hao et al., 2025). Human egocentric video (Grauman et al., 2022; Yang et al., 2025b; Hoque et al., 2026; Zheng et al., 2026) offers a scalable source of experience for developing these capabilities: it records how an actor observes objects, interacts with them, and changes their states over time. Recent robot foundation models (Llontop & Vadrevu, 2026; Robotics, 2026; Luo et al., 2026; Reward AI Team, 2026) use human egocentric video for training; Dyna-2 scales this source to over one million hours. This trend motivates evaluating whether models understand the spatial relations, action sequences, and goals underlying human activities—perceptual and reasoning prerequisites for embodied action, though transfer to robot control remains a separate question. Yet the actor's egocentric view is inherently incomplete: limited coverage and occlusions can hide objects and state changes needed to understand an activity and plan its next steps.

Synchronized exocentric views provide complementary evidence about the same unfolding activity (Sigurdsson et al., 2018; Jia et al., 2020; Rai et al., 2021; Sener et al., 2022; Han et al., 2021). During cooking, for example, the egocentric view may capture detailed hand–object interactions while an external camera reveals an ingredient outside the actor's field of view or a state change obscured by

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.

their hands. Ego-Exo4D (Grauman et al., 2024) makes it possible to study this complementarity in real human activities. Exploiting it requires more than recognizing the same object from different angles: models must track correspondences and state changes over time and select information relevant to the actor's next decision. This raises a central question: *Can MLLMs align spatiotemporal semantics across egocentric and exocentric views, integrate complementary evidence, and use it to make more reliable decisions from the actor's perspective?*

Existing benchmarks evaluate MLLMs on embodied capabilities such as spatial perception (Yang et al., 2025a), planning (Qiu et al., 2026), and embodied cognition (Dang et al., 2025; Luo et al., 2025; Majumdar et al., 2024), but generally do not examine how complementary egocentric and exocentric observations inform decisions. Multi-view benchmarks broaden this scope: MMSI-Bench (Yang et al., 2026) evaluates spatial reasoning across images, while EgoExoBench (He et al., 2026) assesses semantic alignment, viewpoint association, and temporal reasoning across first- and third-person videos. MV-RoboBench (Feng et al., 2025) further connects multi-view spatial understanding with execution-related reasoning in robotic manipulation, primarily using image-based questions. Despite this progress, it remains unclear whether models can use synchronized cross-view evidence from real human activities to support action-oriented understanding, planning, and fine-grained trajectory prediction. Table 1 compares these benchmarks.

To fill this gap, we introduce MultiView-EB, a comprehensive benchmark for cross-view spatiotemporal understanding and planning in real human activities. Built from synchronized ego–exo recordings (Grauman et al., 2024; Cao et al., 2026) through a semi-automatic annotation pipeline with human verification, it comprises eight tasks across three capability dimensions: cross-view perception, task planning, and fine-grained trajectory prediction. The benchmark contains 2,497 question–answer pairs from 658 synchronized ego–exo takes across nine activity domains. Outputs span multiple choice and spatial formats, scored with task-specific metrics for semantic correctness and spatial precision.

We conduct extensive experiments on thirteen general-purpose MLLMs and eight embodied-brain models; most still struggle to establish cross-view correspondences, align temporal dynamics, and selectively use exocentric evidence for egocentric understanding and planning. Failure analysis traces the key limitations and informs directions for future model development. These findings highlight that additional visual inputs alone do not ensure the integrated understanding required for embodied decisions. To investigate how these capabilities can be improved, we construct MV-EmData-38K, on which fine-tuning improves multi-view embodied task performance.

Our contributions are summarized as follows:

- We introduce MultiView-EB, a comprehensive benchmark that evaluates the ability of MLLMs to integrate spatial and temporal information across egocentric and exocentric views to support egocentric embodied tasks in real human activities. It comprises 2,497 question–answer pairs spanning eight tasks across three capability dimensions.
- We develop an automated annotation pipeline that efficiently annotates task-specific metadata and generates question–answer pairs, with subsequent human verification to further improve dataset quality and reliability.
- We conduct extensive experiments on thirteen general-purpose and eight embodied-brain models, revealing persistent difficulties in aligning semantic and temporal information across viewpoints and using exocentric evidence for egocentric understanding and planning. Further analysis of failure modes identifies key limitations and provides insights for future model development.

2 Related Work

We review the two lines of work closest to MultiView-EB; Table 1 contrasts representative benchmarks on observation setting, capability coverage, and scale.

Egocentric embodied understanding. Benchmarks built on large egocentric corpora (Grauman et al., 2022; Yang et al., 2025b) have progressively widened the capabilities tested on the wearer's stream: long-form video understanding and embodied evaluation (Mangalam et al., 2023; Cheng

Table 1: Comparison with representative embodied and egocentric benchmarks on observation setting, capability coverage, and scale. ✓ = capability evaluated; ○ = weaker form (association or anticipation without coordinates, or selection among candidates); × = absent.

	Observation setting				Capability and answer format			
Benchmark	**Domain**	**Views**	**Modality**	**Source**	**Cross-view grounding**	**Planning**	**Trajectory prediction**	**#QA**
VSI-Bench (Yang et al., 2025a)	Human	Ego	Video	ScanNet etc.	×	×	×	5,000+
EgoPlan-Bench2 (Qiu et al., 2026)	Human	Ego	Video	Ego4D	×	✓	×	1,321
ECBench (Dang et al., 2025)	Human	Ego	Video	HM3D etc.	×	×	×	4,324
EgoExoBench (He et al., 2026)	Human	Ego–exo	Video	Ego-Exo4D+5	○	○	×	7,300
MMSI-Bench (Yang et al., 2026)	Mixed	Multi	Image	multi-source	○	×	×	1,000
MV-RoboBench (Feng et al., 2025)	Robot	Multi	Image	AgiWorld etc.	○	✓	○	1,708
RoboBench (Luo et al., 2025)	Robot	Single/multi	Video	robot videos	×	✓	○	6,092
MultiView-EB (ours)	Human	Ego–exo	Video	Ego-Exo4D+ACE	✓	✓	✓	2,497

et al., 2024), hand–object interaction (Xu et al., 2025), spatial intelligence (Yang et al., 2025a), embodied question answering and cognition (Majumdar et al., 2024; Dang et al., 2025), intent understanding and grounding (Pan et al., 2026; Sun et al., 2025), and planning (Qiu et al., 2026; Kundu et al., 2026). All query a single viewpoint, treating occlusions and limited fields of view as nuisance rather than a gap that complementary observations could fill. MultiView-EB anchors every query to the actor's egocentric view but pairs it with a synchronized exocentric one; its matched ego-only and ego+exo conditions turn the value of the second view into a measurable quantity.

Multi-view reasoning. Spatial-reasoning benchmarks have grown from single images or videos (Chen et al., 2024; Song et al., 2025) to multiple views (Yang et al., 2026; Li et al., 2025; Yeh et al., 2026; Gholami et al., 2025), multiple videos (Peng et al., 2025; Zhu et al., 2025), and paired ego–exo streams probing cross-view semantics, memory, and full-body action understanding (He et al., 2026; Liu et al., 2026; Li et al., 2024). In embodied settings, benchmarks evaluate interactive agents (Yang et al., 2025c), spatially grounded planning and embodied-brain capabilities in manipulation (Jung et al., 2026; Luo et al., 2025), and synchronized multi-camera robot workspaces (Feng et al., 2025). These benchmarks query simulated agents, robots, or unstructured collections, and answers remain options or free text rather than coordinates in a queried view; MultiView-EB instead draws on time-synchronized ego–exo video of real human activities and requires grounded outputs—time-stamped boxes, temporal intervals, and 2D waypoints—scored with geometric metrics.

3 MULTIVIEW-EB

3.1 OVERVIEW

We present MultiView-EB, a benchmark for cross-view spatiotemporal understanding in real-world embodied activities. Built from synchronized ego–exo recordings of daily activities, it comprises 2,497 question–answer pairs across eight tasks in three capability dimensions—*cross-view perception*, *task planning*, and *fine-grained trajectory prediction*—tracing an embodied agent's reasoning from locating relevant entities across views, through deciding what to do next, to predicting where the hands will move (Table 4, Appendix A).

Multi-view information plays two roles across the suite. In the perception and trajectory dimensions, the second view is the *object* of the query: a reference view specifies content and the model grounds it in the target view. In planning, it is *evidence*: samples are built so that the decisive state, such as an ingredient outside the wearer's field of view, is observable only exocentrically, and planning is therefore also evaluated under matched ego-only input (Section 4). Answers are heterogeneous—multiple choice, structured JSON, time-stamped boxes, and 2D waypoints—covering both semantic correctness and spatial precision.

3.2 TASK DEFINITION

Each sample is a set of synchronized views and a typed answer. One or more *reference* views specify the content to perceive or supply decision evidence, and the *target* view—the egocentric camera—is where the answer must be expressed, so every spatial answer is given in the actor's frame of



Figure 1: Representative samples from the tasks of MultiView-EB, grouped by capability dimension. Each sample pairs synchronized reference and target views with a typed question and answer.

reference. For tasks querying the future, the observation is cut at a decision point t_d: the model receives the egocentric history $[t_0, t_d]$ plus task-specific conditioning, while the ground-truth future is never shown. Fig. 1 shows one sample per task.

3.2.1 Cross-view perception

The three correspondence tasks assess whether models can establish cross-view correspondences at the object, interaction, and task levels and ground the relevant content in a target view. In *object correspondence*, an exocentric frame highlights an object, and the model tracks it in a synchronized egocentric clip using time-stamped bounding boxes; the object may leave the egocentric field of view for extended periods. In *HOI correspondence*, an exocentric clip depicts a hand–object interaction, and the model locates it within a longer synchronized egocentric clip, returning the interaction interval, the hand and object tracks, and the object's identity and acting hand. In *task correspondence*, given an exocentric clip of a multi-step activity and a target-view frame, the model identifies the objects in their order of use and grounds them in the target frame.

3.2.2 Task planning

The two planning tasks assess whether models can select appropriate actions with the correct target locations. Given a natural-language goal, egocentric observations up to a decision point, and a synchronized exocentric frame, the model chooses from four candidate plans: *next-step planning* selects the immediate action, *long-horizon planning* a sequence of subsequent actions in order. Because the correct decision may involve a target outside the egocentric field of view, the model must incorporate exocentric observations into its plan.

3.2.3 Fine-grained trajectory prediction

The three tasks probe grounding and predicting hand motion in the target view under different conditioning. In *trajectory grounding*, the model maps a manipulation demonstrated in an exocentric video to the corresponding hand trajectories in the image plane of egocentric keyframes. *Goal-conditioned prediction* infers future hand trajectories from the action's before and after states in the reference view and a pre-action observation in the target view; *instruction-conditioned prediction* anticipates them from current ego–exo observations and a natural-language instruction, without seeing the outcome. All three require 2D waypoints for both hands.

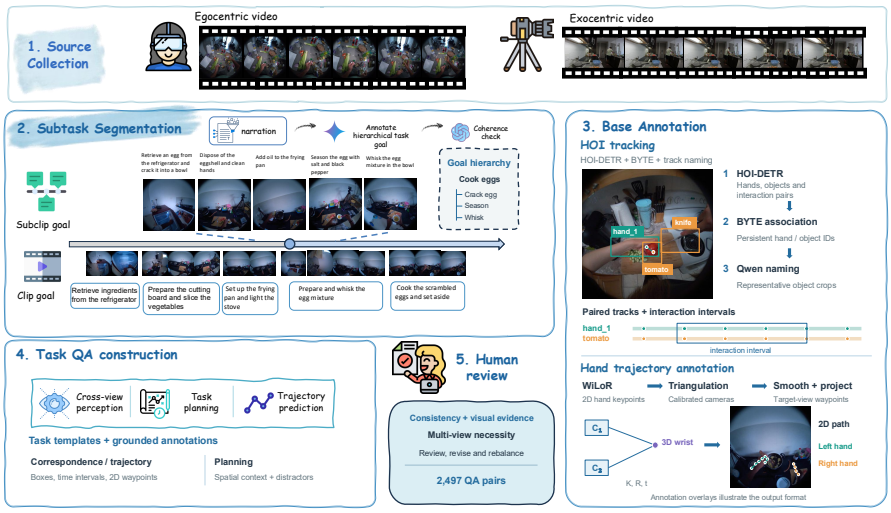


Figure 2: Construction pipeline of MultiView-EB: recordings are segmented (phase 1), annotated with geometric ground truth (phase 2), and assembled into task-specific questions (phase 3), then pass multi-view necessity verification, structured audits, and human review before release.

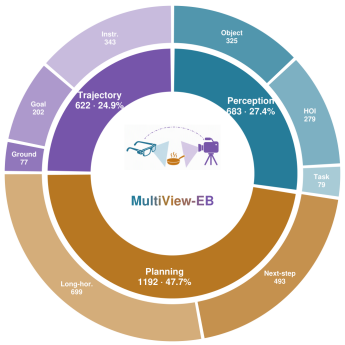
3.3 Benchmark Construction

To evaluate the three capability dimensions, we designed a multi-stage, semi-automatic annotation pipeline with human verification at every gate, which distills open-source multi-view recordings of human activities into high-quality question–answer pairs. We first describe the two source datasets and then the three pipeline phases and their quality control.

MultiView-EB is built on two large-scale collections of synchronized ego–exo recordings. Ego-Exo4D (Grauman et al., 2024) captures skilled human activities across eight scenarios by 740 participants, totaling 1,286 hours; each take pairs the wearer's egocentric camera with multiple synchronized exocentric cameras and carries ordered key steps, dense multi-view action descriptions, and 3D body and hand keypoints with per-camera projections. ACE-Data-0 (Cao et al., 2026) records 150 hours of daily-living interactions by 50 participants across 200 task categories, with time-aligned ego–exo video and human, hand, and object 6D-pose annotations.

Construction proceeds in three phases (Fig. 2). Phase 1 performs hierarchical goal extraction and segmentation: Gemini-3.8-flash parses the narrations accompanying each recording into a hierarchy of goals and segments the video accordingly, and GPT-5.6 checks coherence, filtering idle and task-irrelevant steps. Phase 2 produces the geometric base annotations on two tracks. Native signals are reused wherever the source datasets ship them—cross-view object identities and action narrations in Ego-Exo4D, motion-capture hand–object geometry in ACE-Data-0—and a self-built detection–tracking–naming stack fills the gaps: HOI-DETR (Darkhalil et al., 2026) detects hands, objects, and their interaction pairs; a BYTE tracker (Zhang et al., 2022) associates detections into temporally consistent tracks; Qwen3.8-flash names each object track; and hand trajectories are projected into the target view from 3D wrist keypoints, reconstructed by WiLoR (Potamias et al., 2024) detection and multi-view triangulation where native hand annotation is missing. Phase 3 assembles the question–answer pairs: correspondence and trajectory tasks are instantiated from task-specific templates over the phase-2 annotations, while for planning, Qwen3.8-flash augments each action step with its spatial location and Gemini-3.8-flash generates the distractors.

To ensure reliability, we apply four quality gates: (1) *consistency verification*, a multi-model judge check of every question–answer pair against the underlying annotations, discarding ambiguous or inconsistent items; (2) *vision-grounded filtering*, discarding questions answerable from text alone; (3) a *multi-view necessity check*, verifying for planning questions that the object state deciding the correct answer is invisible in the egocentric observation yet visible in the exocentric one; and (4) *human review*, in which trained annotators discard misaligned items, revise wording and grounding issues, and rebalance option positions. In a stratified audit of 250 questions, 92% satisfied all crite-



Task	#QA	E4D	ACE
Cross-view perception			
Object correspondence	325	200	125
HOI correspondence	279	–	279
Task correspondence	79	–	79
Task planning			
Next-step planning	493	493	–
Long-horizon planning	699	699	–
Fine-grained trajectory prediction			
Trajectory grounding	77	–	77
Prediction (goal)	202	–	202
Prediction (instruction)	343	–	343
Total (8 tasks)	**2,497**	**1,392**	**1,105**

Figure 3: Statistics of MultiView-EB. Left: the 2,497 question–answer pairs by capability dimension (inner ring) and task (outer ring). Right: per-task counts by source dataset (E4D = Ego-Exo4D).

ria and the remainder were corrected rather than removed; overall, 39,509 candidates reduce to the released 2,497 questions, whose distribution over the three capability dimensions, eight tasks, and two source datasets is detailed in Fig. 3.

4 Experiments

4.1 Evaluation Setup

We evaluate thirteen general-purpose MLLMs—seven open-source (Qwen3.5 (Qwen Team, 2026a), Qwen3.6 (Qwen Team, 2026b), Qwen3.8-flash (Qwen Team, 2026c), and InternVL3.5 (Wang et al., 2025), spanning 4B–397B) and six proprietary (GPT-6-astra (OpenAI, 2026b), GPT-6-sol (OpenAI, 2026c), GPT-5.6-sol (OpenAI, 2026a), Claude-Sonnet-5 (Anthropic, 2026), Gemini-3.1-pro and Gemini-3.8-flash (Google, 2025))—together with eight embodied-brain models (Li et al., 2026; Tan et al., 2026; Wang et al., 2026; Liang et al., 2026; Agarwal et al., 2026; Lin et al., 2025; Hao et al., 2025); all names and sizes appear in Table 2. All models run zero-shot with identical prompts across input conditions. The correspondence and trajectory tasks are inherently cross-view; the planning tasks are additionally evaluated under matched ego-only input (Section 4.3), and Table 2 reports the ego+exo condition throughout. Multiple-choice answers are scored by exact match; spatial outputs are parsed into boxes and waypoints and mapped into the target view's pixel space.

4.2 Main Results

No model comes close to solving MultiView-EB. GPT-6-astra leads every task—object correspondence (0.66 mIoU / 0.88 vIoU@0.5), HOI correspondence (0.83 tIoU), task correspondence (0.54 box F1), next- and long-horizon planning (0.582 / 0.720), and all three trajectory conditions (DFD 0.06–0.37)—yet even it misses 28% of long-horizon questions. GPT-6-sol follows, second on every correspondence task and long-horizon planning (0.639); GPT-5.6-sol is second on grounded (DFD 0.116) and instruction-conditioned (0.393) prediction, and Gemini-3.1-pro leads the remaining proprietary models.

Spatial precision, not language ability, is the bottleneck. Models that plan well often fail to ground: Gemini-3.8-flash answers 51.4% of long-horizon questions correctly yet achieves 0.000 vIoU@0.5 on object correspondence, and the InternVL3.5 family shows the same dissociation at up to 38B parameters (0.42 long-horizon accuracy, at most 0.021 mIoU). Within HOI correspondence, temporal localization (tIoU up to 0.83) is far stronger than the hand–object tracks (hand and object IoU at most 0.55 and 0.58): models find *when* an interaction happens but not *where* it is. Small open models trail on every task, and the trajectory conditions order themselves by the strength of the visual conditioning: the best DFD is 0.06 for grounded, 0.31 for goal-conditioned, and 0.37 for instruction-conditioned prediction—the weaker the visual conditioning, the harder the task.

Embodied-brain models do not transfer to MultiView-EB. All eight score 0.000 vIoU@0.5 on object correspondence (best mIoU 0.079, Hy-Embodied-VLM-1.0), and several fail at the protocol level, answering in free-form prose or emitting degenerate trajectory outputs. Isolated bright spots do exist

Table 2: Main results on MultiView-EB under the ego+exo condition; per-task sample counts in Fig. 3. Object: mIoU / vIoU@0.5; HOI: tIoU / hand / object IoU; task: F1 / edit distance (↓); trajectory: discrete Fréchet distance (DFD, ↓). Best in **bold**, second underlined. MV-Emb (ours): Qwen3.5-4B fine-tuned on MV-EmData-38K.

	Cross-view perception			*Task planning*		*Trajectory prediction*		
Model	Obj. mIoU / vIoU	HOI tIoU / hnd / obj	Task F1 / ED ↓	Next Acc.	Long Acc.	Grd. DFD ↓	Goal DFD ↓	Instr. DFD ↓
Open-source general								
Qwen3.5-4B	0.092 / 0.037	0.491 / 0.093 / 0.079	0.012 / 1.06	0.296	0.398	0.643	0.623	0.737
Qwen3.6-35B-A3B	0.122 / 0.040	0.378 / 0.021 / 0.018	0.029 / 1.11	0.310	0.480	0.487	0.440	0.502
Qwen3.5-397B-A17B	0.086 / 0.022	0.493 / 0.017 / 0.019	0.020 / 1.28	0.325	0.461	0.455	0.565	0.630
Qwen3.8-flash	0.331 / 0.289	0.382 / 0.126 / 0.149	0.379 / 1.09	0.379	0.442	0.415	0.433	0.559
InternVL3.5-4B	0.007 / 0.000	0.363 / 0.002 / 0.003	0.000 / 1.25	0.276	0.349	0.887	0.564	0.672
InternVL3.5-8B	0.021 / 0.000	0.234 / 0.001 / 0.004	0.012 / 1.35	0.276	0.405	0.836	0.685	0.816
InternVL3.5-38B	0.016 / 0.000	0.333 / 0.000 / 0.001	0.008 / 1.52	0.282	0.419	0.915	0.701	0.784
Proprietary general								
GPT-5.6-sol	0.482 / 0.529	0.662 / 0.356 / 0.374	0.382 / 1.36	0.456	0.594	0.116	0.356	0.393
GPT-6-astra	**0.660 / 0.880**	**0.825 / 0.545 / 0.581**	**0.539** / 1.61	**0.582**	**0.720**	**0.063**	**0.306**	**0.373**
GPT-6-sol	0.556 / 0.686	0.706 / 0.504 / 0.526	0.478 / 1.46	0.499	0.639	0.146	0.371	0.427
Gemini-3.1-pro	0.096 / 0.049	0.402 / 0.101 / 0.125	0.102 / 1.08	0.472	0.595	0.312	0.515	0.539
Gemini-3.8-flash	0.022 / 0.000	0.420 / 0.030 / 0.034	0.005 / 1.18	0.381	0.514	0.557	0.464	0.577
Claude-Sonnet-5	0.159 / 0.065	0.509 / 0.120 / 0.168	0.018 / 1.21	0.322	0.475	0.299	0.502	0.545
Embodied-brain models								
RynnBrain 1.1-9B	0.049 / 0.000	0.389 / 0.019 / 0.018	0.012 / 1.37	0.333	0.398	0.801	0.740	0.694
RynnBrain 1.1-122B-A10B	0.000 / 0.000	0.002 / 0.000 / 0.000	0.113 / **0.99**	0.327	0.539	0.639	0.468	0.598
RoboBrain 2.5-8B	0.023 / 0.000	0.242 / 0.021 / 0.018	0.005 / 6.11	0.300	0.409	0.600	0.875	0.758
Hy-Embodied-VLM-1.0	0.079 / 0.000	0.521 / 0.066 / 0.083	0.006 / 1.04	0.254	0.366	0.528	0.352	0.444
RxBrain-6B	0.006 / 0.000	0.006 / 0.000 / 0.000	0.000 / 1.03	0.045	0.156	1.414	1.371	1.346
Cosmos3-Nano	0.036 / 0.000	0.412 / 0.045 / 0.049	0.007 / 1.98	0.288	0.428	0.587	0.611	0.585
PhysBrain 1.5-8B	0.017 / 0.000	0.000 / 0.000 / 0.000	0.000 / 6.22	0.308	0.528	1.414	0.556	0.601
MiMo-Embodied-7B	0.035 / 0.000	0.338 / 0.014 / 0.019	0.053 / 2.21	0.280	0.429	1.414	0.417	0.727
Qwen3.5-4B-MV-Emb (ours)	0.508 / 0.594	0.740 / 0.458 / 0.496	0.443 / 1.11	0.545	0.602	0.512	0.361	0.412

relative to general models: RynnBrain 1.1-122B-A10B answers 0.539 of long-horizon questions correctly, above Claude-Sonnet-5, Gemini-3.8-flash, and every open-source general model, and Hy-Embodied-VLM-1.0 attains a goal DFD (0.352) second only to GPT-6-astra. Yet neither model exceeds 0.113 on any spatial-correspondence metric, so each is competitive only on the single task closest to its training distribution—wherever competence exists, it is narrow.

Qwen3.5-4B-MV-Emb, Qwen3.5-4B fine-tuned on the MV-EmData-38K corpus (Appendix G), improves substantially over its zero-shot base on all eight tasks, restoring the cross-view grounding that embodied-brain models lack; Section 4.5 examines how the three capabilities rise together.

4.3 Effect of Exocentric Input

To evaluate whether exocentric observations improve egocentric planning, we conduct a controlled viewpoint ablation on the planning tasks—the only ones with a well-posed ego-only variant. Table 3 reports both planning tasks under ego-only and ego+exo input for eight general-purpose models, with paired McNemar exact tests; its *Act.* and *Loc.*|*Act.* columns decompose the ego+exo answers and are discussed in Section 4.4.

The exocentric evidence frame is largely unused. Adding it improves next-step accuracy by $+2.1$ points on average and long-horizon by only $+0.5$, and only four of the sixteen model–task pairs improve significantly; for GPT-6-astra, the strongest planner, the frame even costs 3.0 points of long-horizon accuracy. Yet the frame is chosen to expose exactly the state changes that decide the answer and are invisible in the egocentric view (Section 3.2.2); the small gains therefore reflect a failure to exploit complementary evidence, not an absence of useful signal.

4.4 Failure Analysis

Planning. Every planning question is factorially constructed—the correct action and location, one distractor sharing the action at a wrong location, two action distractors—so each answer decomposes

Table 3: Planning analysis: input conditions and action–location decomposition (n=493 next-step, n=699 long-horizon). *Ego* = egocentric clip; *Ego+Exo* adds the exocentric frame at the decision point (the condition of Table 2); Δ = Ego+Exo − Ego; *McNemar exact $p < 0.05$. *Act.* = chosen option carries the correct action; *Loc.*$|A$ = exact option given correct action; options are factorially constructed (Section 4.4), with chance baselines in the first row.

	Next-step planning					*Long-horizon planning*				
Model	Ego	Ego+Exo	Δ	Act.	Loc.|A	Ego	Ego+Exo	Δ	Act.	Loc.|A
Chance	0.25	0.25	0.00	0.50	0.50	0.25	0.25	0.00	0.50	0.50
Open-source general										
Qwen3.5-4B	0.262	0.296	+0.034*	0.39	0.76	0.388	0.398	+0.010	0.43	0.92
Qwen3.6-35B-A3B	0.270	0.310	+0.041*	0.39	0.81	0.452	0.481	+0.029*	0.50	0.97
Qwen3.8-flash	0.385	0.379	−0.006	0.51	0.75	0.445	0.442	−0.003	0.51	0.87
Proprietary general										
GPT-5.6-sol	0.436	0.456	+0.020	0.55	0.86	0.588	0.594	+0.006	0.63	0.98
GPT-6-astra	0.535	0.582	+0.047*	0.68	0.90	0.750	0.720	−0.030	0.78	0.96
Gemini-3.1-pro	0.434	0.473	+0.039	0.54	0.88	0.594	0.595	+0.001	0.61	0.98
Gemini-3.8-flash	0.404	0.381	−0.022	0.52	0.74	0.488	0.514	+0.026	0.53	0.97
Claude-Sonnet-5	0.304	0.323	+0.018	0.41	0.78	0.476	0.475	−0.001	0.50	0.95

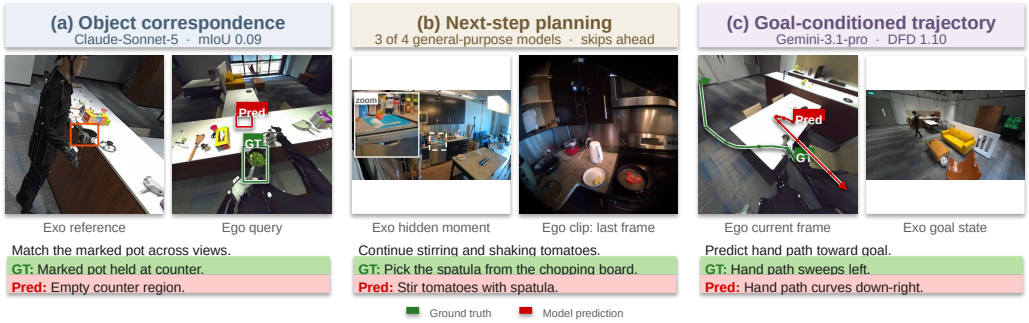


Figure 4: Failure cases (green: ground truth; red: model prediction). **(a)** Object correspondence: Claude-Sonnet-5 binds an empty counter for the pot marked in the exocentric reference. **(b)** Next-step planning: the spatula is still on the chopping board, yet three of four general-purpose models stir with it, skipping the pickup (dashed box: decisive state). **(c)** Goal-conditioned prediction: Gemini-3.1-pro curves the hand path away from the goal-directed ground truth.

into an action decision and a location decision, both with a 0.5 chance baseline (Table 3). Action selection is the binding constraint: next-step action accuracy spans 0.39–0.68 against location-given-action of 0.74–0.90. Planning fails at reading the activity's state and progress from video, not at discriminating spatial options.

Across GPT-6-astra, GPT-6-sol, Gemini-3.1-pro, and Claude-Sonnet-5, 83–97% of planning failures are action errors, and manual inspection of 125 of them finds one dominant mode: *progress miscalibration*, 91%—the right action proposed at the wrong point in the task. The miscalibration is directional and flips with the horizon: next-step errors lag, with 64% repeating a completed step, usually the last one visible in the egocentric clip, whereas long-horizon errors lead, with 55% skipping prerequisite micro-operations. Figure 4(b) shows a consensus case: the exocentric frame shows the spatula still lying on the chopping board, yet three of the four evaluated models choose to stir with it, skipping the prerequisite pickup.

Correspondence. Object-correspondence failure is a binding failure, not a tracking failure (Figure 4(a)): GPT-6-astra binds and tracks the correct object in 94% of samples, whereas Claude-Sonnet-5 and Qwen3.6-35B-A3B bind in only 11–14% yet track nearly as precisely when they do (conditional mIoU 0.37–0.42). A probe on InternVL3.5-38B separates the deficits: 0.66 IoU localizing from an object name in a single frame, 0.38 transferring a marked exocentric reference, but 0.016 mIoU on the full multi-frame task—single-image localization is intact, single-frame transfer partial, and binding collapses once multiple frames compete.

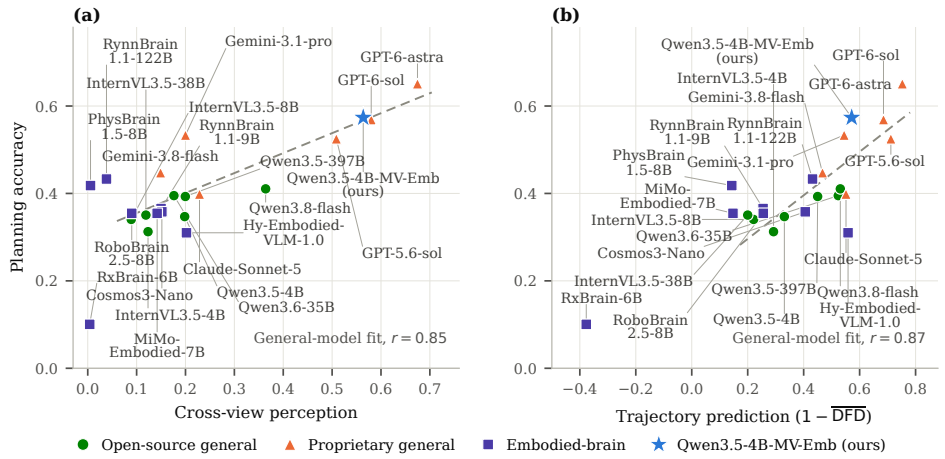


Figure 5: Capability-dimension relations per model (ego+exo; Table 2). (a) Cross-view perception (mean of object mIoU, HOI tIoU, task box F1) vs. planning accuracy (mean of next-step and long-horizon). (b) Trajectory prediction (1− mean DFD) vs. planning accuracy. Dashed lines are OLS fits of the general-purpose models ($r = 0.85$ and $r = 0.87$); the MV-Emb fine-tune (star) is excluded.

Trajectory. The same qualitative misreading appears in goal-conditioned trajectory prediction: predictions are fluent but geometrically wrong—Gemini-3.1-pro curves the hand path down-right where the ground truth sweeps left toward the goal state (Figure 4(c)).

4.5 Cross-Dimension Correlation

Figure 5(a) relates correspondence perception to planning. Among the general-purpose models, the two correlate strongly (Pearson $r = 0.85$, $n = 13$). The embodied-brain models decouple the dimensions: with near-zero perception (0.00–0.20) they plan at 0.10–0.43, bracketing the fitted floor. Planning is the dimension most porous to language priors; object-level correspondence is where it shows.

Figure 5(b) repeats the comparison with fine-grained trajectory prediction, and the embodied-brain models rejoin the trend: prediction correlates with planning about as strongly as perception does ($r = 0.87$, $n = 13$), and the relationship extends across model classes (pooled over all twenty-one zero-shot models, $r = 0.83$; four of the eight embodied-brain models fall within ± 0.07 of the general-model fit). Predicting where the hands will move is the spatial capability that travels with planning; object-level correspondence is where embodied models collapse. The MV-Emb fine-tune (star) shows the capability profile is learnable: fine-tuning Qwen3.5-4B on MV-EmData lifts it from the lower end of the zero-shot field onto the general-model trend of panel (a), with all three dimensions rising together (perception $0.20 \rightarrow 0.56$, planning $0.35 \rightarrow 0.57$, trajectory $0.33 \rightarrow 0.57$) rather than trading off.

5 Conclusion

We present MultiView-EB, a benchmark for cross-view spatiotemporal understanding in real-world embodied activities: eight tasks across three capability dimensions. Evaluating thirteen general-purpose MLLMs and eight embodied-brain models shows that spatial precision, not language ability, is the binding constraint and that embodied-brain models largely fail to transfer. The viewpoint ablation quantifies the central claim: an exocentric frame that exposes the state deciding the answer lifts next-step/long-horizon planning accuracy by only 2.1/0.5 points on average—access to more views does not by itself produce cross-view understanding. Failure analysis traces the deficit to progress miscalibration and cross-view binding, and fine-tuning on MV-EmData-38K improves all eight tasks. MultiView-EB and MV-EmData provide a systematic testbed for MLLMs that selectively exploit complementary evidence for the actor's next decision.

AI USE STATEMENT

We acknowledge the use of large language models as writing assistants for grammar and style, and for generating synthetic datasets. Beyond these uses, LLMs are not employed in the core research methodology, experimental design, data analysis, or generation of research findings presented in this paper. All textual content and LLM-generated data have been reviewed and verified by the authors.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No new human-subject data was collected: MultiView-EB is derived entirely from two publicly available ego–exo datasets, Ego-Exo4D (Grauman et al., 2024) and ACE-Data-0 (Cao et al., 2026), each recorded by its own team with informed participant consent and used here in accordance with the terms of its public release. Our contribution consists of question–answer annotations, prompts, and evaluation code built on top of these releases; we collect no new recordings and introduce no personally identifiable information beyond what the source datasets already distribute. Human involvement was limited to annotation review, conducted by trained annotators following written annotation guides (Appendix C).

MultiView-EB is intended solely for evaluating and improving embodied perception and planning models. Its scenario coverage inherits the domains of the two source datasets—skilled activities and daily living—and conclusions drawn from it should not be generalized beyond these settings. We are not aware of harmful dual-use applications of the released artifacts, and no experiments were conducted that raise privacy or security concerns. We are committed to transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. Section 3.3 and Appendix C document the benchmark construction in full: the task suite specification with inputs, answer schemas, and metric suites; per-task construction details; the four quality gates; and qualitative examples with intermediate pipeline outputs. All prompts used for construction and evaluation are quoted verbatim in Appendix H. Appendix D details the benchmark composition, including the assignment of source takes between the benchmark and the MV-EmData corpus, whose construction and full fine-tuning configuration are described in Appendix G. Appendix E specifies the evaluated models and how they are accessed, the inference and decoding configuration, output parsing and scoring conventions, and the definitions of all evaluation metrics, and Appendix F documents the error-analysis protocol. Together, these materials cover the complete path from source recordings to every number reported in the paper.

Both source datasets used in this paper, Ego-Exo4D (Grauman et al., 2024) and ACE-Data-0 (Cao et al., 2026), are publicly available, enabling consistent re-derivation and evaluation. The MultiView-EB benchmark, the complete evaluation harness, and the MV-EmData fine-tuning corpus have been made available in an anonymous repository to facilitate replication and verification.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Anthropic. Claude Sonnet 5 system card. System card, June 2026. URL `https://www.anthropic.com/claude-sonnet-5-system-card`.

Yukang Cao, Haozhe Xie, Beichen Wen, Runmao Yao, Yinghao Liu, Yue Huang, Zhichao Liao, Yunxiang Wang, Haiheng Liu, Xingshun Tian, et al. Ace-data-0: Human-centric ambient capture as embodied data engine. *arXiv preprint arXiv:2607.28625*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *arXiv preprint arXiv:1702.05448*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain 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*, pp. 14455–14465, 2024.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. VidEgoThink: Assessing egocentric video understanding capabilities for embodied AI. *arXiv preprint arXiv:2410.11623*, 2024.

Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. *arXiv preprint arXiv:2105.11731*, 2021.

Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24593–24602. IEEE, 2025.

Ahmad Darkhalil, Dima Damen, and David Fouhey. Improving and evaluating hand-object interaction detection. *arXiv preprint arXiv:2606.17384*, 2026.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *arXiv preprint arXiv:2211.03726*, 2022.

Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. *arXiv preprint arXiv:2510.19400*, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. *arXiv preprint arXiv:1705.02101*, 2017.

Mohsen Gholami, Ahmad Rezaei, et al. Spatial reasoning with vision-language models in egocentric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025. Introduces Ego3D-Bench.

Google. A new era of intelligence with Gemini 3. Official announcement, November 2025. URL `https://blog.google/products/gemini/gemini-3`.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 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.

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.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *arXiv preprint arXiv:1803.10892*, 2018.

Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2021.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first-and third-person view video understanding in mllms. *Advances in Neural Information Processing Systems*, 38, 2026.

Ryan Hoque, Peide Huang, David J. Yoon, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.

Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision*, pp. 767–786. Springer, 2020.

Sehun Jung, HyunJee Song, Dong-Hee Kim, Reuben Tan, Jianfeng Gao, Yong Jae Lee, and Donghyun Kim. Spatially grounded long-horizon task planning in the wild. *arXiv preprint arXiv:2603.13433*, 2026. Introduces GroundedPlanBench.

Kaustav Kundu, Ritvik Shrivastava, Maxim Arap, et al. Plan, watch, recover: A benchmark and architectures for proactive procedural assistance. *arXiv preprint arXiv:2606.04970*, 2026. Introduces Pro2Bench.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

Kehan Li, Bohan Hou, Minghao Zhu, Tianyi Zhang, Zesen Cheng, Zhikai Wang, Sicong Leng, Xin Li, Xiao Lin, Biying Yao, et al. Rynnbrain 1.1: Towards more capable and generalizable embodied foundation model. *arXiv preprint arXiv:2607.17977*, 2026.

Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. EgoExo-Fitness: Towards egocentric and exocentric full-body action understanding. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2406.08877.

Haotian Liang, Mingkang Chen, Yufei Huang, Yuchun Guo, Xiaomeng Zhu, Xiangli Shi, Kaixuan Wang, Yunxuan Mao, Weijie Zhou, Ling Chen, et al. Rxbrain: Embodied cognition foundation model with joint language-visual reasoning and imagination. *arXiv preprint arXiv:2607.14187*, 2026.

Xiaopeng Lin, Shijie Lian, Bin Yu, Ruoqi Yang, Zhaolong Shen, Changti Wu, Yuzhuo Miao, Yurun Jin, Yukun Shi, Jiyan He, et al. Physbrain: Human egocentric data as a bridge from vision language models to physical intelligence. *arXiv preprint arXiv:2512.16793*, 2025.

Ruiping Liu, Junwei Zheng, Yufan Chen, Di Wen, Shaofang Quan, Chengzhi Wu, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. EgoExoMem: Cross-view memory reasoning over synchronized egocentric and exocentric videos. *arXiv preprint arXiv:2605.18734*, 2026.

Edith Llontop and Kalyan Vadrevu. Nvidia isaac gr00t n1. 7: Open reasoning vla model for humanoid robots. *Hugging Face Blog, April*, 2026.

Hao Luo, Wanpeng Zhang, Yicheng Feng, et al. Being-h0.7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Xiangju Mi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, et al. Robobench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. *arXiv preprint arXiv:2510.17801*, 2025.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. Official model announcement, July 2026a. URL https://openai.com/index/gpt-5-6.

OpenAI. GPT-6 Astra: A new generation of intelligence. Official model announcement, September 2026b. URL https://openai.com/index/gpt-6-astra.

OpenAI. Introducing GPT-6 Sol and luna. Official model announcement, September 2026c. URL https://openai.com/index/introducing-gpt-6-sol-and-luna.

Ye Pan, Chi Kit Wong, Yuanhuiyi Lyu, Hanqian Li, Chenfei Liao, Jiahao Huo, Lutao Jiang, Zixin Zhang, Jiacheng Chen, Yuqian Fu, and Xu Zheng. EgoIntent: A pre-outcome micro-step benchmark for understanding what, why, and next. *arXiv preprint arXiv:2603.12147*, 2026.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, et al. MVU-Eval: Towards multi-video understanding evaluation for multimodal LLMs. *arXiv preprint arXiv:2511.07250*, 2025.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *International Journal of Computer Vision*, 134(5):222, 2026.

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

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. Technical report, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency. Technical report, August 2026c. URL https://qwen.ai/blog?id=qwen3.8-flash-next.

Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11179–11188. IEEE, 2021.

Reward AI Team. Om-1: Frontier robot intelligence, learned firsthand from humans. Reward AI Blog, 2026. URL https://www.rewardai.com/blog/OM-1/.

Dyna Robotics. Dyna-2: A 1-million-hour scaling law for world-action models. *Dyna Robotics, Official Research Report*, 2026.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21064–21074. IEEE, 2022.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7396–7404. IEEE, 2018.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780. IEEE, 2025.

Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. Introduces the EgoIntention dataset; arXiv:2504.13621.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, et al. Robobrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Ziyi Wang, Xumin Yu, Yongming Rao, Yonggen Ling, Yunheng Li, Oran Wang, Mingqi Gao, Yuchen Zhou, Yves Liang, Zuyan Liu, et al. Hy-embodied-vlm-1.0: Efficient physical-world agents. *arXiv preprint arXiv:2607.12894*, 2026.

Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, and Qin Jin. Do egocentric video-language models truly understand hand-object interactions? In *International Conference on Learning Representations (ICLR)*, 2025. Introduces EgoHOIBench; arXiv:2405.17719.

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

Jingkang Yang, Shuai Liu, Hongming Guo, et al. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025b.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *International Conference on Machine Learning (ICML)*, 2025c. arXiv:2502.09560.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12000–12008, 2026.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv preprint arXiv:2505.08548*, 2025.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. *arXiv preprint arXiv:2001.06891*, 2020.

Ruijie Zheng, Dantong Niu, Yuqi Xie, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Nannan Zhu, Yonghao Dong, Teng Wang, et al. CVBench: Benchmarking cross-video synergies for complex multimodal reasoning. *arXiv preprint arXiv:2508.19542*, 2025.

A Appendix Overview

This appendix supplements the main text with the following material.

- **Appendix B, Comparison with Existing Benchmarks**: introduces each benchmark summarized in Table 1 and details how MultiView-EB differs in observation setting, capability coverage, and answer format, including the criteria behind the capability columns of Table 1.
- **Appendix C, Task Suite and Benchmark Construction**: the complete specification of the eight tasks—inputs, answer schemas, and metric suites—together with per-task construction details, the quality-control gates applied at each stage of the pipeline, qualitative examples from both source datasets, and intermediate pipeline outputs.
- **Appendix D, Additional Statistics**: composition by domain and source, clip and annotation distributions, answer-type and option balance, and the assignment of source takes between the benchmark and the fine-tuning corpus.
- **Appendix E, Experimental Setup**: the evaluated models and how they are accessed, inference and decoding configuration, output parsing and scoring conventions, and the definitions of all evaluation metrics.
- **Appendix F, Per-task Error Analysis**: an error taxonomy for each task, quantified over sampled failures and illustrated with model reasoning cases.
- **Appendix G, Fine-tuning on MV-EmData**: the construction and composition of MV-EmData, the training configuration, and full fine-tuning results.
- **Appendix H, Complete Prompt Packages**: the prompts used for benchmark construction and evaluation, quoted verbatim.

B Comparison with Existing Benchmarks

This section expands Table 1: we introduce the compared benchmarks grouped by line of work, detail how MultiView-EB differs from each group, and state the criteria behind the observation-setting and capability columns. Full specifications of the eight MultiView-EB tasks appear in Section 3.2 and Appendix C.

B.1 Compared Benchmarks

Egocentric and embodied single-view benchmarks. **VSI-Bench** (Yang et al., 2025a) evaluates visual-spatial intelligence on first-person walkthrough videos of indoor scenes: over 5,000 question–answer pairs from 288 videos of ScanNet, ScanNet++, and ARKitScenes, across configurational, measurement-estimation, and spatiotemporal tasks answered as multiple choice or numeric values (the latter scored with mean relative accuracy). **EgoPlan-Bench2** (Qiu et al., 2026) assesses task planning in real-world scenarios from the first-person perspective: 1,321 multiple-choice questions from Ego4D (Grauman et al., 2022) spanning four major domains and 24 scenarios, each presenting a goal, a history clip, and the current observation, with distractors drawn from other steps of the same task. **ECBench** (Dang et al., 2025) targets holistic embodied cognition on egocentric RGB-D video: 4,324 question–answer pairs over 30 cognitive dimensions, mixing open-ended and multiple-choice formats, built from HM3D navigation streams, ScanNet and MultiScan scans, and self-collected depth-camera footage of counterintuitive and dynamic scenes.

Multi-view and robot benchmarks. **MMSI-Bench** (Yang et al., 2026) probes spatial reasoning across multiple images with 1,000 multiple-choice questions curated from more than 120,000 real-world images pooled from eight sources (indoor scans, egocentric video, street scenes); every non-multi-step question pairs exactly two images and cannot be answered from either image alone. **MV-RoboBench** (Feng et al., 2025) evaluates multi-view spatial reasoning in robotic manipulation: 1,708 manually annotated five-choice questions over eight subtasks in two categories (spatial understanding and robotic execution), built from 980 AgiWorld and BridgeV2 episodes in single- and dual-arm settings, with synchronized multi-view image pairs rather than video. **RoboBench** (Luo et al., 2025) evaluates MLLMs as embodied manipulation brains with 6,092 question–answer pairs over 4,336 unique items covering five dimensions, 14 capabilities, and 25 tasks, combining open-source robot recordings with in-house collection across single-arm, dual-arm, mobile, and humanoid embodiments; inputs are sparsely sampled frames, and open-ended long-horizon plans are scored by an MLLM world simulator.

Ego–exo benchmark. **EgoExoBench** (He et al., 2026), the closest in observation setting, is the first benchmark dedicated to ego–exo video understanding: over 7,300 four-way multiple-choice questions across 11 subtasks in three dimensions—Ego–Exo Relation, View Transition, and Temporal Reasoning—aggregating six ego–exo datasets, four with synchronized multi-view recordings and two asynchronous demonstration–follower pairs.

B.2 Differentiation from MultiView-EB

We detail the differences in input synchronization, answer form, and trajectory output that Table 1 does not encode, per group below.

Single-view egocentric benchmarks. VSI-Bench, EgoPlan-Bench2, and ECBench observe an activity from one camera, so every question inherits that camera's blind spots. MultiView-EB instead builds every sample on synchronized ego–exo video of the same real activity—658 takes from Ego-Exo4D and ACE-Data-0 (Cao et al., 2026)—with the second view playing the two roles of Section 3.1: object of the query (perception, trajectory) or decision evidence visible only exocentrically (planning). EgoPlan-Bench2 is the closest predecessor: our planning tasks adopt its goal-conditioned multiple-choice format and add the exocentric evidence its single-view setting cannot provide; VSI-Bench's numeric estimates stay within one view, whereas MultiView-EB requires typed spatial outputs in a designated target view; ECBench evaluates embodied cognition without goal-conditioned planning or motion prediction.

Multi-view and robot benchmarks. These benchmarks provide multiple views but without video (MMSI-Bench, MV-RoboBench) or without human activities (MV-RoboBench, RoboBench), and their spatially grounded questions reduce to candidate selection. MultiView-EB differs in both observation and output: synchronized ego–exo video of real human activities rather than image pairs of manipulation scenes; time-stamped boxes and temporal intervals in the target view rather than view-level matching (MV-RoboBench is closest in spirit but only matches objects across synchronized camera pairs); and directly predicted 2D waypoints for both hands rather than selection among candidates (the trajectory selection of MV-RoboBench and the affordance questions of RoboBench).

Ego–exo benchmark. EgoExoBench matches MultiView-EB in pairing egocentric and exocentric video of human activities and shares Ego-Exo4D as a source, but all its questions are multiple-choice: cross-view understanding is tested as semantic association and temporal reasoning as anticipation and ordering, with no spatial coordinates in any answer and no trajectory prediction. MultiView-EB instead requires heterogeneous typed answers evaluated with spatially sensitive metrics, predicts hand trajectories directly, and constructs planning samples whose decisive state is visible only exocentrically; moreover, all 658 MultiView-EB takes are synchronized recordings, whereas two of EgoExoBench's six sources pair asynchronous demonstrations with follower videos.

B.3 Column Keys of Table 1

All rows of Table 1 describe real-world observations, and all question counts follow the original papers.

Observation-setting columns. *Domain* records who acts: *Human* (egocentric or ego–exo recordings), *Robot* (robot manipulation recordings), *Mixed* (multi-domain image collections). *Views* records how many viewpoints a question presents and how they relate: *Ego* (single first-person camera), *Ego–exo* (wearer's camera plus synchronized external cameras), *Multi* (robot-mounted and external cameras, or multi-image sets without synchronized capture), *Single/multi* (RoboBench, which mixes both). *Modality* records the visual input—video clips versus static image pairs and sets (RoboBench: frames sparsely sampled from robot videos). *Source* names the underlying data collection: MultiView-EB builds on Ego-Exo4D and ACE-Data-0, and EgoExoBench on Ego-Exo4D and five further ego–exo datasets.

Capability columns. The three capability columns correspond to the capability dimensions of MultiView-EB (Section 3.1); ✓ indicates the capability is evaluated, ∘ a weaker form, and × absence. *Cross-view grounding*: ✓ requires localizing, with spatial coordinates, content specified in one view within another view—in MultiView-EB, time-stamped boxes and temporal intervals in a target view; ∘ marks cross-view semantic association or matching without spatial coordinates (EgoExoBench's relation and view-transition subtasks, MMSI-Bench's cross-image spatial relations, MV-RoboBench's cross-view matching); × marks single-view inputs. *Planning*: ✓ marks goal-conditioned selection of the next action or an action sequence (EgoPlan-Bench2, MV-RoboBench, and RoboBench, which additionally judges open-ended plans with an MLLM world simulator); ∘ marks EgoExoBench's anticipation and ordering subtasks; ECBench is × (egocentric question answering rather than goal-conditioned planning), and VSI-Bench is × because its route-plan task is grouped by its authors under configurational understanding. *Trajectory prediction*: ✓ requires producing motion trajectories as output—in MultiView-EB, 2D waypoints for both hands in the target view; ∘ marks selection among trajectory or grasp candidates (MV-RoboBench, RoboBench); × marks absence.

C Task Suite and Benchmark Construction

This section gives the complete, reproducible specification of MultiView-EB: sample schema (Sec. C.2), source data and signals (Sec. C.3), the three-phase pipeline and its actual detection, tracking, and naming components (Sec. C.4), per-task details for all eight tasks (Sec. C.5), and quality control with the construction funnel (Sec. C.6).

C.1 Task Suite

Table 4 summarizes the eight tasks (view roles as in Sec. 3.2). Figures 9–16 show complete, evaluated samples of every task—two per task, from both sources across seven domains, none shared with the main-text examples—placed alongside the per-task details below; Figs. 6–8 show real intermediate pipeline outputs.

C.2 Unified Sample Schema

All tasks share one sample schema, so a single evaluation harness consumes the whole benchmark. A sample is a record with four groups (Table 5); media live in a per-sample directory referenced by the manifest.

Leakage prevention. For every task that queries the future, the observation is cut strictly at a decision point t_d and the ground-truth future never enters the input; MCQ option order is shuffled with an RNG keyed by the sample ID, so the golden position is uniform by construction (Table 7).

C.3 Source Data and Signals

Ego-Exo4D. From Ego-Exo4D (Grauman et al., 2024) we consume four annotation families, and a take enters a task's candidate pool only when the annotations that task requires are all present: (i) ordered *key steps*; (ii) *narrations* with per-view visibility flags (`ego_visible`, plus the best exo camera when exocentrically visible); (iii) *relations*, per-camera object masks and boxes in which one physical object carries one identifier across all cameras; and (iv) *ego-pose*, 3D body and hand

Table 4: The task suite of MultiView-EB. Every sample pairs views with complementary roles—reference views specify the content to perceive or provide decision evidence, and the target view is where the answer is required—together with an instruction and task context. The Metrics column lists the primary reported metric; full metric suites and protocols follow the corresponding evaluation literature.

Task	Input	Answer	Metrics
Cross-view perception			
Object correspondence	exo frames marking one object; ego clip	time-stamped box track in ego	mIoU, vIoU@0.5
HOI correspondence	exo interaction clip; longer ego clip	time range; hand–object track	tIoU, hand/obj IoU
Task correspondence	exo multi-step clip; one target frame	ordered boxes; null if invisible	box F1, seq. edit
Task planning			
Next-step planning	goal; ego clip to decision point; exo frame at that point	MCQ: action	accuracy
Long-horizon planning	goal; ego clip to decision point; exo frame at that point	MCQ: three-action sequence	accuracy
Fine-grained trajectory prediction			
Trajectory grounding	reference action clip; target keyframe	left/right-hand 2D waypoints	DFD
Prediction (goal)	reference before/after; target before	2D waypoints (future)	DFD
Prediction (instruction)	ego and exo observation frames; instruction	2D waypoints (future)	DFD

keypoints with per-camera projections, plus per-frame camera poses (Aria intrinsics/extrinsics) and GoPro fisheye calibrations. Planning additionally requires the per-frame camera poses, which drive hidden-event mining and decision-point anchoring, and the self-built trajectory annotation requires the native-resolution export (1408-pixel Aria and 4K exo); all other tasks run on the standard downscaled export. In total, 576 takes contribute, spanning all eight source scenarios.

ACE-Data-0. From ACE-Data-0 (Cao et al., 2026) we consume 82 takes of room-scale daily-living activity. The decisive signals are the motion-capture ones: SMPL-X body parameters (wrist joints by forward kinematics), per-object 6D poses, and time-aligned multi-view video (one HOMIE egocentric plus several ZED exocentric cameras). The mocap frame grid maps linearly onto every video timeline, so a 3D hand–object contact episode can be projected into any camera—this anchors the ground truth of all five ACE tasks. The dataset's rendered hand overlays serve as an independent cross-check of our projection chain (Sec. C.6).

Ground-truth anchors. Together, cross-camera object identities (relations) and per-camera keypoint projections (ego-pose; mocap plus calibrated projection in ACE-Data-0) state, for any moment and view, whether the decisive content is visible—the basis of the multi-view necessity check in Sec. C.6.

C.4 Construction Pipeline

Construction proceeds in three phases (Fig. 2 in the main text; real intermediate outputs in Figs. 6–8). Phase 2 is dual-track: native annotations are reused wherever the source datasets ship the needed geometry; the self-built detection–tracking–naming stack runs only where native signals are missing.

C.4.1 Phase 1: segmentation and hierarchical goal extraction

For each Ego-Exo4D take, Gemini-3.8-flash parses the timestamped narrations into a three-level goal tree—*overall goal*, *clips* (subgoals), *sub-clips*—with timestamps copied from the narrations (Fig. 6); a GPT-5.6 coherence pass and a per-clip purposefulness filter remove idle, task-irrelevant, and duplicate narrations. The tree supplies the goals of the planning tasks; a goal is eligible when it contains 4–20 kept actions.

The planning tasks are anchored at *hidden events*: moments whose decisive state is invisible egocentrically but visible exocentrically. Candidates are mined by a camera-pose signal: at each boundary of the phase-1 segmentation, the net rotation of the egocentric (Aria) camera over the transition

Table 5: Unified sample schema and ground-truth payloads. Times are seconds relative to the start of the target clip; boxes are $[x_1, y_1, x_2, y_2]$ in target-view pixels; waypoints are (u, v) in target-view pixels. Every sample is validated at write time: views must carry a role and media type, MCQ choices must be distinct with an in-range golden index, and the provenance group must record both construction phases.

Group	Field	Content
identity	task_type	task identifier, e.g. P1_object_centric_correspondence
	sample_id	unique; also the per-sample media directory name
	take_uid / take_name	source take (Ego-Exo4D UID or ACE-Data-0 take)
	scenario / split	activity scenario; train or val
input	views	list of views; each with role (reference | target | context), camera ID, derived viewpoint (egocentric/exocentric), media_type (frame | frames | clip), media file(s), time range, and optional prompt_marks (e.g. the marked box drawn on an exo reference frame)
	instruction	the natural-language question shown to the model
	reference_semantics	task/action/intent text for prediction and planning tasks; empty for pure-perception tasks
answer	answer_type	bbox_seq | point_seq | waypoint_2d | waypoint_3d | mcq | open_text
	value	the ground-truth payload (per-task rows below)
	choices / golden_choice_idx	MCQ only: option texts and the index of the correct option
	distractor_meta	MCQ only: per-choice provenance and design tags (Sec. C.4)
provenance	segmentation_source	phase-1 source, e.g. keystep, goal_hierarchy, hidden_events_camera_pose, time_window, contact_episode_3d
	base_annotation_source	phase-2 source, e.g. relations, atomic_descriptions, ace_mocap, fused_hoi+sam3, hoi_detr@5fps+bytetrack+vlm, selfbuilt_trajectory(wilor_gtfree)
	needs_review / filter_review	review flags and the filtering verdicts of Sec. C.6
	extra	task-specific fields (QC metrics, clip windows, gates, option design)
Ground-truth payload (answer.value) by task — task | answer_type | payload		
Object corr.	bbox_seq	$[\{t_i, \text{bbox}_i, \text{label}\}]$ over the target clip (E4D: 1 s grid; ACE: 0.5 s grid covering the full clip)
HOI corr.	open_text	$\{\text{time_range}:[t_s, t_e]$, hoi_track: $[\{t_i$, object box + label, hand box + label$\}]\}$ on a 0.5 s grid
Task corr.	bbox_seq	$[\{k, \text{bbox}_k \mid \text{null}, \text{label}_k\}]$, k=1…K objects in order of use; null when invisible in the target frame
Next-step / long-horizon	mcq	four options; each option is an action phrase with a parenthesized spatial location (long-horizon: a →-joined three-action sequence)
Trajectory grounding	waypoint_2d	$[\{\text{frame}_j$, per-frame $[\{(u, v)$, left/right hand$\}]\}]$ for each target keyframe
Pred. (goal / instr.)	waypoint_2d	$[\{t_i, (u, v)_i, \text{left_hand}\|\text{right_hand}\}]$, t_i on a 0.2 s grid after the decision point



Figure 6: Real intermediate output of phase-1 hierarchical goal extraction (Ego-Exo4D take `sfu_cooking028_2`, *Cooking an Omelet*, 183 narrations). Gemini-3.8-flash parses the timestamped narrations into a three-level goal tree—overall goal, clips, sub-clips—with every time span copied from the narrations; each row carries its real span as a bar on the take's 404 s timeline (one color per clip, lighter bars for sub-clips). The three sampled ego frames (colored borders keyed to the clip spans; ▼ markers on the axis) each fall inside a marked sub-clip span.

into the next step is measured, and boundaries whose rotation exceeds a 35° threshold become candidates—the wearer turned or relocated to begin the next step, so what the next step acts on typically lies outside the pre-turn egocentric field of view. For goal-tree takes, whose subgoal boundaries are narration-derived, the decision point is the frame of maximum camera rotation within a ± 5 s window around the boundary—the actual head-turn instant—and the exocentric reference camera is chosen geometrically, as the camera best framing what the wearer turns toward. The prediction target is the step following the change; the relevance and multi-view necessity gates of Sec. C.6 then verify, per candidate, that the decisive state is ego-invisible and exo-visible.

C.4.2 PHASE 2: BASE ANNOTATION

Phase 2 produces, per take and view, the geometric base annotations from which answers are derived; Table 6 lists, per task, which track produced the released ground truth.

Table 6: Phase-2 ground-truth sources of the released questions. "Native" rows reuse source-dataset annotations; "self-built" rows run our detection–tracking–naming stack. The E4D self-built outputs additionally constitute the E4D part of the fine-tuning corpus (Appendix G).

Task (source)	**Track**	**GT derivation**
Object corr. (E4D)	native `relations`	the marked object's cross-camera track ID yields its boxes in the ego camera; target frames sampled at 1 fps
Object corr. (ACE)	native mocap	object 6D pose projected into views; visibility gate (40° viewing angle / 0.30 m); boxes on a 0.5 s grid over the full clip
HOI corr. (ACE)	native mocap	3D hand–object contact episodes (distance $\leq$ 0.08 m) define the interaction window; hand blocks and object poses projected into the ego view
Task corr. (ACE)	native mocap	ordered contact episodes give the object-use order; per-object boxes projected into the target frame
Planning (E4D)	native narrations	golden action = the next kept narration after t_d; its spatial location is grounded by Qwen3.8-flash on per-action frames
Trajectory (ACE)	native mocap	SMPL-X wrist joints (FK) per frame, projected into the target view through calibrated camera parameters
HOI tracking (E4D, self-built)	self-built	HOI-DETR detects hands, object roles, and interaction pairs; two BYTE trackers (Zhang et al., 2022) associate detections (hands by velocity-predicted center distance, objects by inflated IoU); Qwen3.8-flash names each object track from representative crops and classifies hand fragments by side, followed by take-level side validation
Hand traj. (E4D, self-built)	self-built	takes with ego-pose hand annotation: native 3D hand keypoints, RTS-smoothed; takes without: WiLoR (YOLO hand detector + MANO reconstructor) 2D wrists in the exo 4K views, multi-view triangulated and RTS-smoothed; wrist positions projected into the target view

Self-built HOI tracking (Ego-Exo4D). The self-built stack runs a detect–track–name cascade (Fig. 7). HOI-DETR runs at 5 fps and detects hands, two object roles, and pairwise interaction links; because the roles are roles rather than identities—the same object swaps roles across frames—all object detections are tracked jointly in one BYTE tracker (Zhang et al., 2022) and hands in a second, with hands associated by velocity-predicted center distance (tiny, fast ego-view hand boxes yield zero IoU between frames) and objects by inflated IoU. Qwen3.8-flash then names each object track from its most representative crops (top score-by-area frames, majority vote) and classifies hand fragments left/right, followed by a take-level geometric side check; a dense alternative replaces detection boxes with SAM3 (Carion et al., 2026) mask tracks where pixel-accurate identity is needed.

Self-built hand trajectories (Ego-Exo4D). For takes without native hand keypoints, wrist trajectories are reconstructed without ground truth (Fig. 8): WiLoR (YOLO hand detector + MANO reconstructor) regresses 21 2D keypoints per hand in each exocentric 4K view, wrists are triangulated across at least three views (with an Aria-based fallback), and world-frame trajectories are smoothed by robust outlier rejection plus a constant-velocity RTS smoother. Takes with ego-pose hand annotation skip detection and project the native 3D keypoints with the same smoothing. In both cases the target-view ground truth is the world-frame wrist projected through the target view's calibrated parameters (undistorted 512-pixel Aria pinhole domain for egocentric targets; GoPro fisheye intrinsics with static extrinsics for exocentric ones); the artifact covers 517 validation-split takes.

ACE refinement chain. On the ACE side, all five tasks share one mocap-derived base—3D contact episodes, SMPL-X wrist joints, and object 6D poses, projected into the relevant cameras—refined by a chain that keeps the mocap geometry authoritative: Qwen3.8-flash refines projected

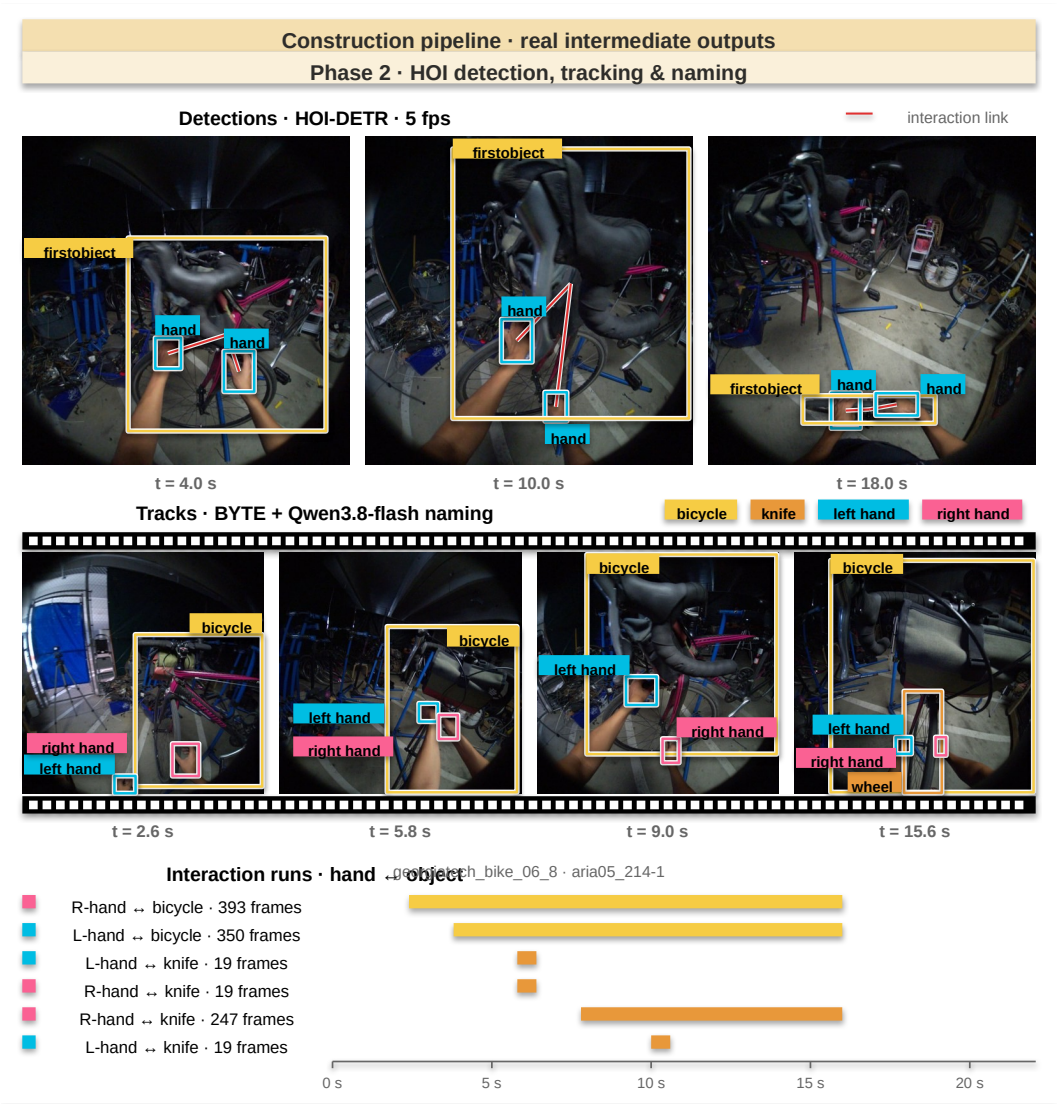


Figure 7: Real intermediate outputs of the self-built HOI stack on Ego-Exo4D take `georgiatech_bike_06_8` (ego camera `aria05_214-1`). Top: raw HOI-DETR detections at 5 fps—class-colored boxes with red interaction links between interacting hand–object pairs. Middle: the same video after the two BYTE trackers and Qwen3.8-flash naming—one color per identity across the strip (left/right hand, bicycle, knife; boxes drawn verbatim from the released track artifact). Bottom: the hand–object interaction runs over the take (bar color = object, swatch = hand side).

hand and object boxes on sampled frames, accepted only under IoU, shift, and containment gates; for object correspondence, the exocentric reference box is re-localized by GPT-6-astra on the clean frame and adjudicated per sample by human review (Sec. C.6).

C.4.3 Phase 3: Question Assembly

Correspondence and trajectory questions are instantiated from task-specific templates over the phase-2 base annotations (per-task parameters in Sec. C.5); planning questions are multiple-choice with a strict distractor discipline:

- *Real-text pools.* Every option surface is a real kept narration of the same take or a real location phrase; unfillable candidates are dropped rather than padded.

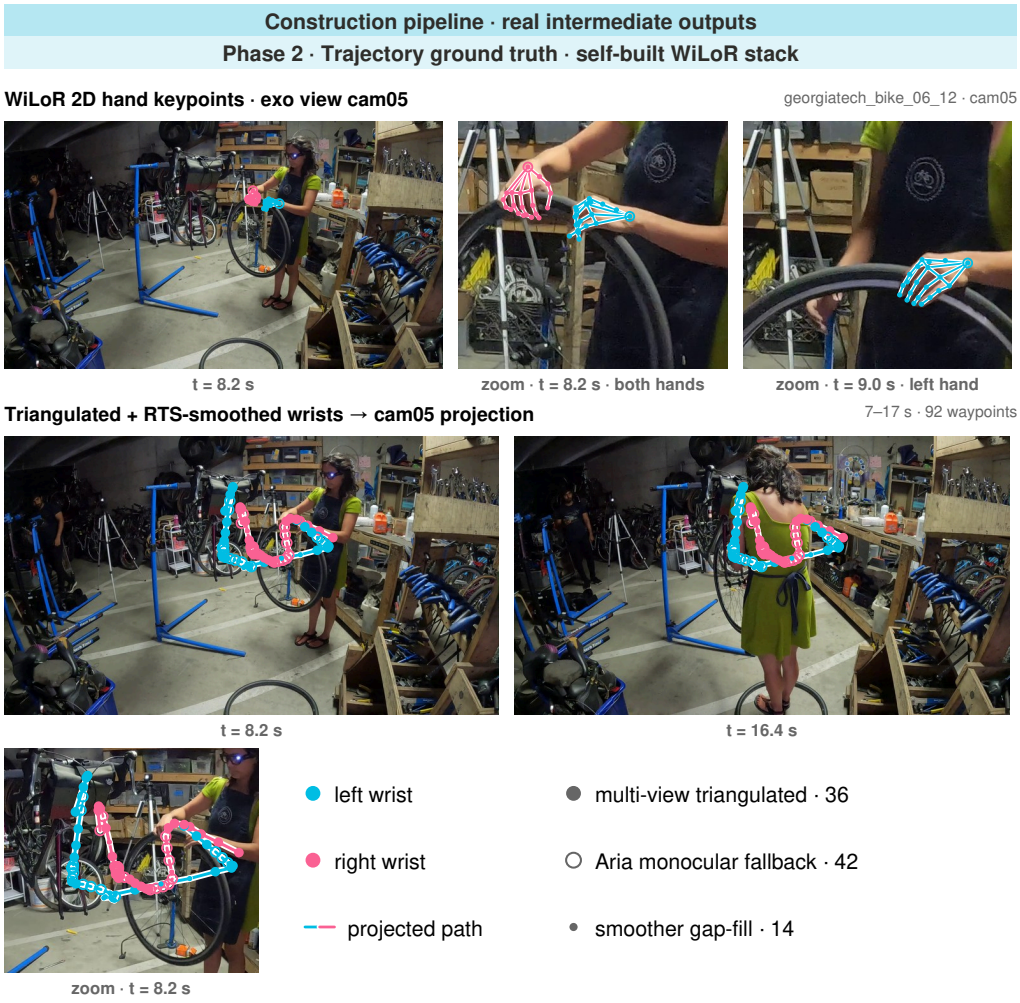


Figure 8: Real intermediate outputs of the self-built trajectory stack on Ego-Exo4D take `georgiatech_bike_06_12`, which has no native hand annotation. Top: WiLoR regresses 21 2D keypoints per hand in the exocentric 4K view `cam05` (left hand cyan, right pink). Bottom: the released GT-free ground truth—2D wrists triangulated across views, RTS-smoothed, and projected back into `cam05` through the pipeline's own GoPro fisheye calibration (10 s window, 92 waypoints on 48 frames; filled dots = multi-view triangulated, hollow = Aria monocular fallback, small = smoother gap-fill).

- *Factorial action×location design.* Each question pairs the golden action with its true location and the same action with an adversarial wrong location—exactly two options share the golden action skeleton, distinguished only by location.
- *Adversarial locations.* The wrong location is not merely different furniture: a decoy is accepted only if a blind three-judge text panel cannot distinguish it from the true location; steps with no gate-passing decoy keep a plain different-location option and are tagged.
- *Temporal-state distractors.* At least one distractor is an already-completed action of the same goal failing a repeatable-action gate—wrong only because of *when* it happens; the rest come from semantically distinct action clusters of the same goal.
- *Deterministic option shuffle.* Option order is shuffled per sample with an RNG keyed by the sample ID (Table 7); duplicate options abort the candidate.

C.5 Per-Task Construction Details

C.5.1 Object correspondence

Input. One exocentric reference frame with a single marked object (vector box), and an egocentric target clip spanning the reference moment: an E4D frame sequence at 1 fps or the ACE clip around the object's contact episode.

Ground truth. E4D: the marked object's relations track ID yields its boxes in the ego camera; ACE: the object's 6D pose is projected into the ego view, a box being emitted only past a viewability gate (viewing angle $\leq 40°$, distance ≤ 0.30 m, at least three in-frame points), the track extending over the full clip (Table 6).

Difficulty design. Samples are selected so that the target object varies substantially in scale across views, may be absent from the egocentric field of view for multiple seconds, and co-occurs with visually similar objects that the reference frame must disambiguate.

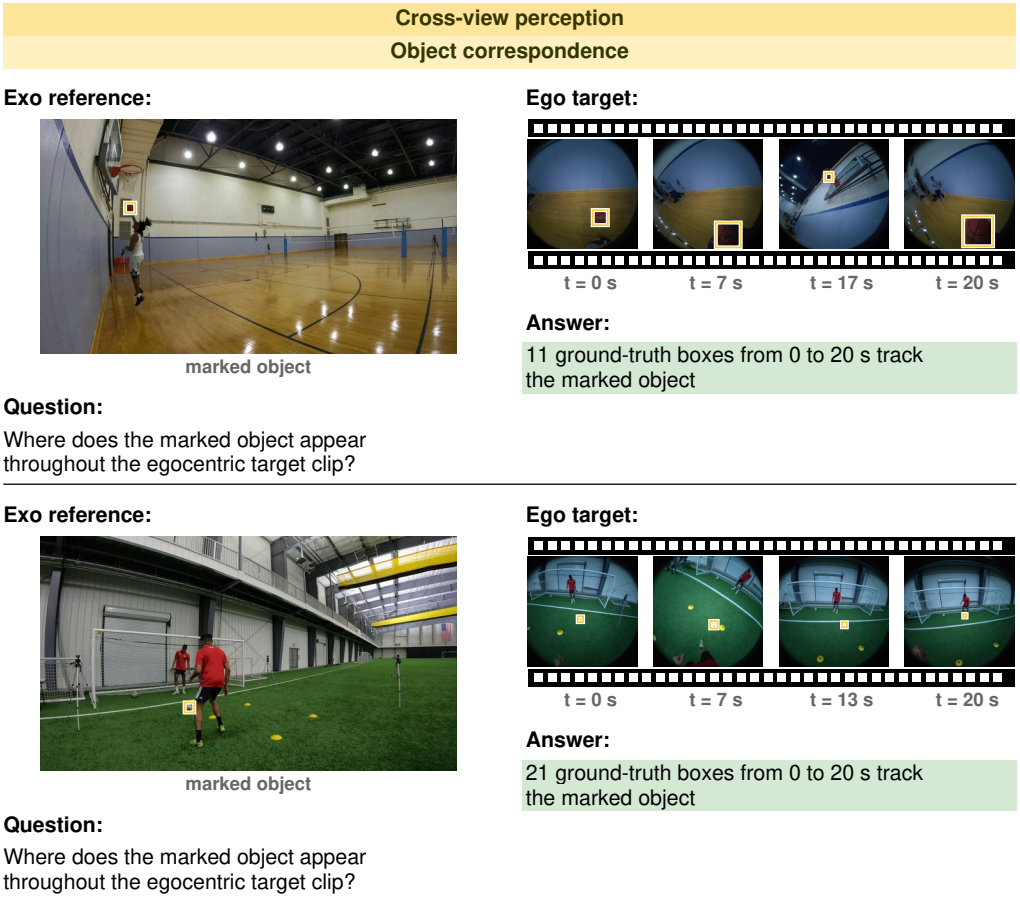


Figure 9: Worked examples of object correspondence (Ego-Exo4D), in the case-figure layout used throughout this appendix: query views and question on the left, answer views and ground truth on the right. Each example shows the marked exocentric reference frame and four frames of the egocentric target clip with relations-derived GT boxes (yellow; times relative to clip start); the shown frames sample the full track (top: basketball, 11 boxes over 0–20 s; bottom: soccer ball, 21 boxes over 0–20 s). GT overlays are vector annotations, not part of the input.

C.5.2 HOI CORRESPONDENCE

Input. An exocentric clip of one hand–object interaction and a longer synchronized egocentric clip that contains it among other interactions.

Ground truth. A 3D mocap contact episode (hand–object distance ≤ 0.08 m) defines the interaction window; the contacting hand block and object pose are projected into the ego view on a 0.5 s grid inside the window (payload in Table 5).

Difficulty design. The longer ego clip contains distractor interactions with other objects and the other hand, so the *referenced* instance must be retrieved; left/right identity and tIoU-scored temporal boundaries add discrimination.

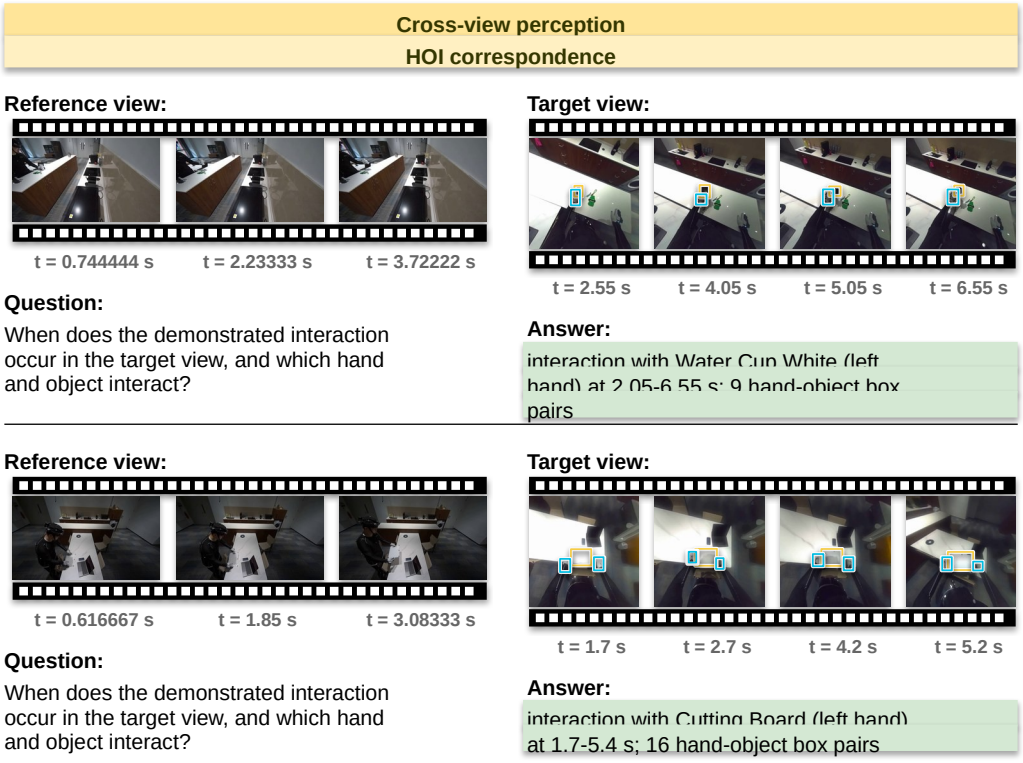


Figure 10: Worked examples of HOI correspondence (ACE-Data-0; views labeled reference/target as the source metadata does not follow the ego/exo distinction). In each example, the reference clip (left, three frames with clip-relative timestamps) depicts the referenced interaction; the longer target clip (right) contains it among distractors. Yellow: GT object boxes; cyan: GT box of the acting hand. Top: water cup, left hand, 2.05–6.55 s, 9 hand–object box pairs; bottom: cutting board, left hand, 1.7–5.4 s, 16 pairs (0.5 s grid). GT overlays are vector annotations, not part of the input.

C.5.3 TASK CORRESPONDENCE

Input. An exocentric clip of a multi-step activity and a single target-view frame.

Ground truth. The ordered contact episodes of the clip give the objects in their order of use; each object's projected box in the target frame is emitted with its order index, or `null` when not visible there.

Difficulty design. Some procedure objects are absent from the target frame by construction (used earlier or elsewhere); the box-F1 / sequence-edit metrics reward exact order and placement.

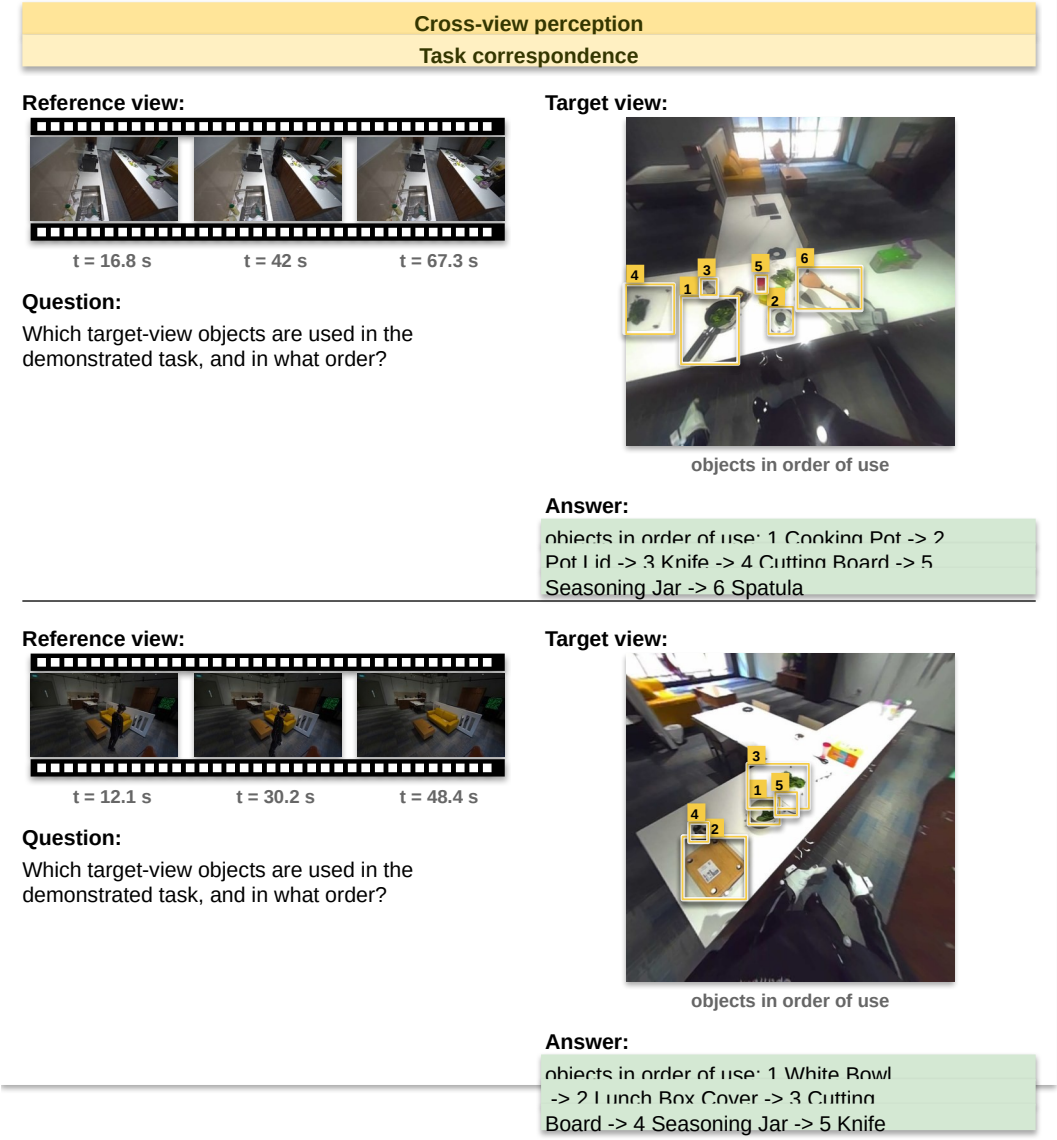


Figure 11: Worked examples of task correspondence (ACE-Data-0). In each example, the reference clip (left, three frames with clip-relative timestamps) shows the multi-step activity; the single target-view frame (right) carries the procedure objects numbered in their order of use (top: 1 Cooking Pot $\rightarrow$ 2 Pot Lid $\rightarrow$ 3 Knife $\rightarrow$ 4 Cutting Board $\rightarrow$ 5 Seasoning Jar $\rightarrow$ 6 Spatula; bottom: 1 White Bowl $\rightarrow$ 2 Lunch Box $\rightarrow$ 3 Lunch Box Cover $\rightarrow$ 4 Cutting Board $\rightarrow$ 5 Seasoning Jar $\rightarrow$ 6 Knife). Numbers encode temporal use order from the ordered contact episodes, not screen position; boxes are annotation-derived GT vectors.

C.5.4 Next-step and long-horizon planning

Input. A natural-language goal (an eligible goal of the phase-1 tree), the egocentric history up to the decision point t_d (at most 300 s), and one exocentric frame at the hidden-event moment; the observation point is set 0.5 s before the golden action, or at the camera-pose anchor when earlier.

Ground truth. The golden answer is the next kept atomic action after t_d, as an action phrase plus its grounded spatial location; long-horizon questions use the next three kept actions, joined in order.

Task planning

Next-step planning

Ego history:



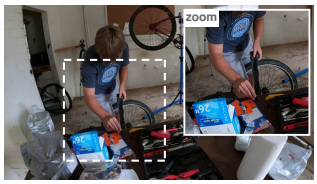
Goal:

Unpack and slightly inflate the new inner tube

Question:

Given the goal and observed history, which action and location should come next?

Exo view:



Options:

A. pick air pump from floor (on the floor beside the blue repair stand)

B. pick air pump from floor (on the floor near the bicycle wheel)

C. cut plastic wrap on tube (on the workbench)

D. place valve cap on table (on the work table near the tool tray)

Ego history:



Goal:

Advance further upward onto orange handholds and footholds

Question:

Given the goal and observed history, which action and location should come next?

Exo view:



Options:

A. place right foot on big orange-colored foothold (on the grey climbing wall)

B. grab big handhold (on the climbing wall)

C. place left foot on crimson foothold (on the climbing wall)

D. place right foot on big orange-colored foothold (on the padded floor mat)

Figure 12: Worked examples of next-step planning (Ego-Exo4D; bike repair top, rock climbing bottom). Each example shows the goal, four frames of the egocentric history up to the decision point (clip-relative timestamps), and the single exocentric frame at the hidden-event moment—the decisive state is ego-invisible and exo-visible (dashed box: zoom on the decisive state). All four options are listed verbatim in their released, deterministically shuffled order, with the ground-truth option highlighted in green. The factorial action×location design is visible: two options share the golden action skeleton and differ only in the location (top: A and B; bottom: A and D). Planning examples carry no spatial GT overlay.

Distractors. The factorial action×location design of Sec. C.4.3: the golden action with an adversarially gated wrong location, at least one already-completed temporal distractor, and same-goal cluster distractors; long-horizon distractors are three-action sequences from the same pools.

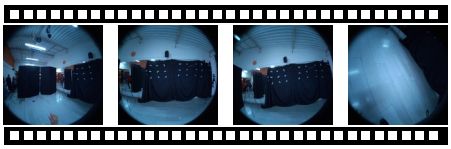
C.5.5 Trajectory grounding

Input. An exocentric reference clip demonstrating the manipulation, and 2–4 egocentric keyframes selected at the largest camera-pose changes inside the annotated window (rotation-threshold selection, coarse to fine until the minimum count is met), each with its own image plane.

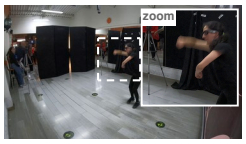
Task planning

Long-horizon planning

Ego history:



Exo view:



exo · now

Goal:

Execute alternating kicks, back steps, and turns

Question:

Given the goal and observed history, which action sequence and locations should follow?

Options:

A. step backward with left leg (in front of the black backdrop) -> step backward with left leg (in front of the black backdrop) -> turn to the left (to the left of the black backdrop)

B. kick up left leg (on the floor near the seated person's feet) -> step backward with left leg (on the floor mat near the orange wall section) -> kick up left leg (on the floor mat in front of the large pink banner)

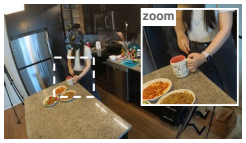
C. turn to the left (in the open doorway on the left) -> kick up right leg (on the light-colored floor in the center of the room) -> step backward while sweeping hands up (in front of the black curtain)

D. kick up left leg (on the floor mat near the black backdrop) -> step backward with left leg (on the floor mat near the tripod) -> kick up left leg (on the floor mat in front of the black curtain)

Ego history:



Exo view:



exo · now

Goal:

Steep tea and add more sugar into the cup

Question:

Given the goal and observed history, which action sequence and locations should follow?

Options:

A. open sugar container lid (on the kitchen island) -> pick sugar container from countertop (on the counter near the stove) -> pour sugar into cup (in the white bowl on the counter)

B. open sugar container lid (on the kitchen counter near the sink) -> pick sugar container from countertop (on the kitchen counter near the sink) -> pour sugar into cup (in the hand holding the mug)

C. place sugar container on countertop (on the counter near the sink) -> hold tea bag string (in the mug held in the left hand) -> stir tea bag in cup (inside the mug held in the left hand)

D. steep tea bag in cup (on the granite kitchen island) -> pick cup from table (on the granite kitchen island) -> drop tea bag string (inside the mug on the granite kitchen island)

Figure 13: Worked examples of long-horizon planning (Ego-Exo4D; salsa dance top, tea preparation bottom). Goal, four timestamped history frames, and the exocentric frame at the hidden-event moment as in Fig. 12 (dashed box: zoom on the decisive state); the answer is a $\rightarrow$-joined three-action sequence. In each example, two options share the golden three-action skeleton and differ only in their locations (top: B and D, gold D; bottom: A and B, gold B), illustrating the factorial design at the sequence level. Options are listed verbatim in their released order.

Ground truth. Per keyframe, the SMPL-X wrist joints of both hands are projected into that keyframe's camera, yielding independent left/right waypoint lists on a fixed grid; waypoints that leave the frame are dropped under the in-frame margin gate.

Difficulty design. Per-keyframe tracks are scored independently; the keyframe-end margin and minimum-segment gates remove degenerate windows.

Fine-grained trajectory prediction

Trajectory grounding

Reference view:

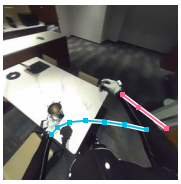


t = 1.32778 s t = 3.98333 s t = 6.63889 s

Question:

Where do the demonstrated left and right wrist paths project in each target keyframe?

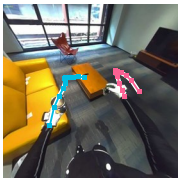
Target view:



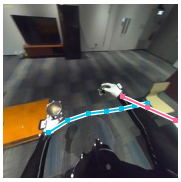
keyframe 1



keyframe 2



keyframe 3



keyframe 4

Answer:

left/right wrist waypoint polylines grounded independently in 4 target keyframes

---

Reference view:



t = 1.17778 s t = 3.53333 s t = 5.88889 s

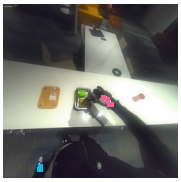
Question:

Where do the demonstrated left and right wrist paths project in each target keyframe?

Target view:



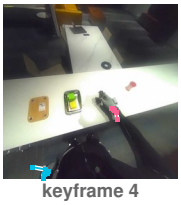
keyframe 1



keyframe 2



keyframe 3



keyframe 4

Answer:

left/right wrist waypoint polylines grounded independently in 4 target keyframes

Figure 14: Worked examples of trajectory grounding (ACE-Data-0). In each example, the reference clip (left, three frames with clip-relative timestamps) demonstrates the manipulation; the four target-view keyframes (right) carry the left/right wrist waypoint polylines (cyan/pink) grounded in each keyframe's own image coordinates. Paths never connect across keyframes; waypoints are projected SMPL-X wrist joints drawn as editable vectors.

C.5.6 GOAL-CONDITIONED TRAJECTORY PREDICTION

Input. Exocentric frames showing the scene before and after the action (the goal state), and the egocentric target frame immediately before the action; an endpoint gate requires the action's endpoint to be visible in the reference view, so the goal is well defined.

Ground truth. The future wrist trajectory from t_d over the action window, projected into the target view on a 0.2 s grid.

Difficulty design. The outcome is observed only exocentrically and only as a state change, so the hand motion that produces it must be inferred in the target view's geometry.

Fine-grained trajectory prediction

Prediction (goal)

Reference view:



before

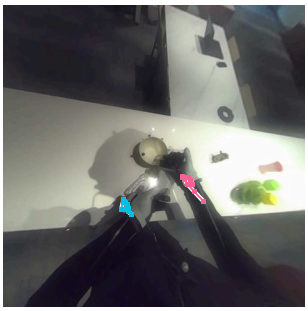


goal state

Question:

What future wrist paths in the target view will reach the demonstrated goal state?

Target view:



now → GT future

Answer:

future left/right wrist waypoint polylines on the target view (28 points)

Reference view:



before

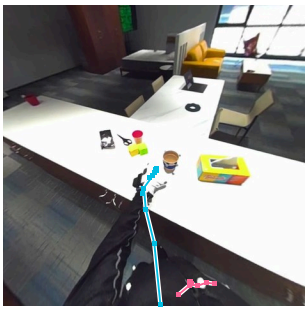


goal state

Question:

What future wrist paths in the target view will reach the demonstrated goal state?

Target view:



now → GT future

Answer:

future left/right wrist waypoint polylines on the target view (21 points)

Figure 15: Worked examples of goal-conditioned trajectory prediction (ACE-Data-0). In each example, the reference view shows the scene before and after the action (left: before, goal state); the target view (right) is the observation immediately before the action, with the annotation-derived future wrist trajectory (cyan left, pink right; 28 and 21 waypoints on the 0.2 s grid) that leads to the goal state. The outcome is observed only in the reference view; the GT path is a vector overlay, not part of the input.

C.5.7 INSTRUCTION-CONDITIONED TRAJECTORY PREDICTION

Input. Egocentric and exocentric observation frames at t_d and a natural-language instruction describing the upcoming manipulation (merged from the source annotations; instructions may reference objects not currently visible).

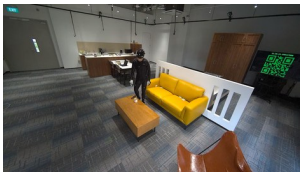
Ground truth. The future wrist trajectory in the target view, on the same grid and projection chain as goal-conditioned prediction.

Difficulty design. No view shows the outcome; every sample passes the instruction-misalignment audit of Sec. C.6, and hand-side mismatches between instruction text and mocap contact are flagged in `provenance`.

Fine-grained trajectory prediction

Prediction (instruction)

Reference view:



now

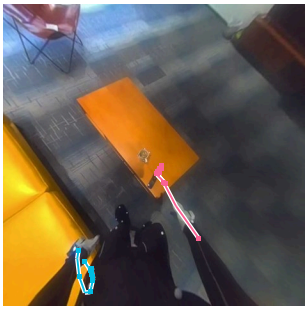
Instruction:

"pick up the empty cup from the orange coffee table"

Question:

What future left and right wrist paths in the target view will fulfill the instruction?

Target view:

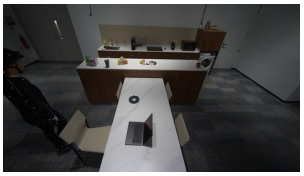


now → GT future

Answer:

future left/right wrist waypoint polylines on the target view (42 points)

Reference view:



now

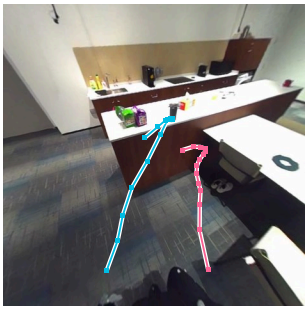
Instruction:

"pick up the metal cup from the white table next to the snack bags"

Question:

What future left and right wrist paths in the target view will fulfill the instruction?

Target view:



now → GT future

Answer:

future left/right wrist waypoint polylines on the target view (29 points)

Figure 16: Worked examples of instruction-conditioned trajectory prediction (ACE-Data-0). In each example, the reference and target views show the multi-view observation at the decision point; the instruction (top: "pick up the empty cup from the orange coffee table"; bottom: "pick up the metal cup from the white table next to the snack bags") together with this observation must determine the future wrist trajectory in the target view (cyan/pink; 42 and 29 waypoints). No view shows the outcome; the GT path is a vector overlay, not part of the input.

C.6 Quality Control and Construction Funnel

Four gate families run between construction and release, followed by targeted ground-truth audits and human review.

(1) Consistency verification. An LLM judge checks every question–answer pair against the underlying annotations and discards or repairs ambiguous or inconsistent items. For planning, two stronger instruments apply: a *privileged-information judge* that verifies facts the examinee cannot see (with positive and negative controls and two fact modes to quantify leakage bias), and a *constructive uniqueness check* in which one model proposes an equally reasonable alternative answer and two blind models verify it—a doubly verified alternative removes the sample.

(2) Vision-grounded filtering. Beyond the adversarial-location design (Sec. C.4.3), a visual-verifiability gate classifies the probe frames of every planning sample by three-model majority vote—`object_visible`, `container_opaque`, `location_only`, or `location_not_visible`—and discards all but the first, removing questions answerable from priors even with the exocentric frame.

(3) Multi-view necessity check. For every planning sample, five ego frames across the history window and the exocentric frame at the hidden-event moment are judged separately by Qwen3.8-flash (strict visibility semantics for ego—the target or named place itself must be clearly visible; looser for exo, accepting partial visibility and hand-held targets); a sample is kept only if the decisive state is ego-invisible and exo-visible.

(4) Human review. Trained annotators review each task's question pool on per-task platforms with the actual media embedded, following written annotation guides; review actions include discarding misaligned items, rewording, regrounding, and rebalancing option positions, and the planning manifests carry `needs_review` flags through release. The instruction-misalignment audit was validated against human ground truth (7/7 agreement on the audited subset).

Targeted ground-truth audits. Beyond the per-sample gates we audit the geometric chains themselves: a QR-to-video synchronization audit compares our projected hand clouds against the dataset's rendered hand overlays, alongside projection-direction, clip-integrity, and granularity checks; a misalignment sweep re-judged every instruction-conditioned sample (removals in Table 7).

Construction funnel. Table 7 reports the per-task funnel from candidates to released questions with the dominant drop reasons; the planning rows count the full planning build (1,486 and 879 takes), from which both the released planning questions and the planning portion of the fine-tuning corpus (Appendix G) derive.

Audit coverage. The stratified audit summarized in the main text (250 questions sampled across the eight tasks and both sources, 92% passing all criteria, the remainder corrected in place) applies the criteria of the review streams above.

D Additional Statistics

This section complements the benchmark statistics (Fig. 3) with the distributions behind the headline counts. Unless stated otherwise, all statistics use all 2,497 released question–answer pairs, with object correspondence pooling its 125 ACE-Data-0 and 200 Ego-Exo4D samples.

D.1 Composition by Source and Domain

Table 8 breaks the eight tasks down by data source and activity domain. The 1,105 ACE-Data-0 pairs (44.3% of the benchmark) all come from its room-scale daily-living environment, the single largest domain; cooking is the largest Ego-Exo4D domain with 931 pairs (37.3%), followed by bike repair (210; 8.4%) and basketball (110; 4.4%), with the remaining five domains contributing 141 pairs combined. The sources are complementary: ACE-side perception and all trajectory tasks are

Table 7: Construction funnel per task, with the golden-option position histograms (A/B/C/D) of the released planning questions. "Candidates" are mined decision points (planning) or constructed samples (all other tasks); drops are the dominant reasons in order. Planning rows count the full planning build (retaining 8,941 and 2,462 samples), of which 493 next-step and 699 long-horizon pairs are released; no histogram deviates more than 2.9 percentage points from the uniform 25% per position. E4D = Ego-Exo4D.

Task	Src.	Cand.	Rel.	Gold pos.	Dominant drops
Next-step planning	E4D	18,983	493	133/117/124/119	no eligible goal (4,986); duplicate golden atomic (3,604); distractor starvation (1,440)
Long-horizon planning	E4D	18,904	699	161/195/186/157	no eligible goal (8,016); distractor starvation (7,289); duplicate golden (1,136)
Object corr.	E4D	442	200	–	one-sample-per-take stratified selection
Object corr.	ACE	149	125	–	unsuitable reference targets in human adjudication (24)
HOI corr.	ACE	279	279	–	—
Task corr.	ACE	79	79	–	—
Traj. grounding	ACE	77	77	–	—
Prediction (goal)	ACE	202	202	–	—
Prediction (instr.)	ACE	394	343	–	instruction misalignment (51)

Table 8: Question counts by task and activity domain (object correspondence pools both sources). Ego-Exo4D takes are assigned to the eight source domains by their take names; the room-scale ACE-Data-0 activities form the ninth domain. "–" marks combinations with no samples.

Task	ACE	Cook.	Bike	Bask.	Health	Dance	Music	Rock	Soccer	Total
Cross-view perception										
Object corr.	125	81	31	49	15	–	22	–	2	325
HOI corr.	279	–	–	–	–	–	–	–	–	279
Task corr.	79	–	–	–	–	–	–	–	–	79
Task planning										
Next-step	–	329	93	23	12	29	–	6	1	493
Long-horizon	–	521	86	38	24	20	1	4	5	699
Fine-grained trajectory prediction										
Traj. grounding	77	–	–	–	–	–	–	–	–	77
Pred. (goal)	202	–	–	–	–	–	–	–	–	202
Pred. (instr.)	343	–	–	–	–	–	–	–	–	343
Total	**1,105**	**931**	**210**	**110**	**51**	**49**	**23**	**10**	**8**	**2,497**

ACE-only, planning is Ego-Exo4D-only, and object correspondence is the only task mixing both (125 vs. 200).

D.2 Input Duration and Ground-Truth Structure

Figure 17 shows the input and ground-truth distributions; Table 9 summarizes them per task. Input duration is measured on the target-view input (egocentric history for planning, target clip for correspondence); image-based tasks carry no duration. Durations span three orders of magnitude: object-correspondence clips are short (median 17.6 s; 12.9 s for the ACE-Data-0 clips and 20.0 s for the Ego-Exo4D frame sequences), HOI-correspondence clips are the shortest (8.6 s), and planning histories are the longest (98.8 s and 111.8 s median for next-step and long-horizon planning, capped at 300 s by construction). Planning histories are bimodal (medians of 48.0 s and 121.8 s for next-

Table 9: Per-task statistics (object correspondence pools both sources). Takes is the number of unique source takes (object correspondence: 63 ACE-Data-0 + 200 Ego-Exo4D). Input duration is the target-view input in seconds, median [min–max]; image-based tasks carry no duration, and planning histories are capped at 300 s. Ground-truth output is median [min–max] over samples: boxes per track (object correspondence), tracked entries and steps per sample (HOI, task correspondence, planning), and waypoints per sample (trajectory tasks); planning gold answers are fixed-length.

Task	#QA	Takes	Answer type	Duration (s)	GT output
Object corr.	325	263	box sequence	17.6 [2.0–20.0]	17 boxes [3–40]
HOI corr.	279	68	open JSON	8.6 [1.4–23.5]	9 entries [3–33]
Task corr.	79	79	box sequence	–	3 steps [1–21]
Next-step	493	312	multiple choice	98.8 [0.8–300.0]	1 step (fixed)
Long-horizon	699	356	multiple choice	111.8 [1.1–300.0]	3 steps (fixed)
Traj. grounding	77	77	waypoints	–	72 waypoints [32–86]
Pred. (goal)	202	66	waypoints	–	30 waypoints [6–62]
Pred. (instr.)	343	70	waypoints	–	32 waypoints [6–62]

step, 61.4 s and 140.1 s for long-horizon), visible as the two-step shape of the planning ECDFs in Fig. 17a. Across the 1,796 clip-based samples the pooled median duration is 30.8 s.

Ground-truth structure varies correspondingly. Object-correspondence tracks contain a median of 17 boxes (20 for ACE-Data-0, sampled across the full clip, and 15 for Ego-Exo4D); HOI-correspondence answers localize interaction windows of median length 4.0 s (0.6–8.0 s) with a median of 9 tracked entries; task-correspondence answers list 3 steps per sample at the median (1–21), 8 of the 286 steps (2.8%) carrying `null` boxes because the object is not visible in the target frame; planning gold answers are fixed-length by design (one action for next-step, three for long-horizon); and trajectory answers contain a median of 72 waypoints per sample for trajectory grounding, spread over 3–4 target keyframes (median 4), and 30 and 32 waypoints for goal- and instruction-conditioned prediction, roughly half per hand (per-hand medians 15.5–16).

Panel (b) of Fig. 17 gives the ground-truth box-area distribution that underlies the target-size analysis of Section 4.4, expressed relative to the target frame ($1{,}024^2$ for ACE-Data-0, $1{,}408^2$ for Ego-Exo4D). Median box area is 0.82% of the frame for ACE-Data-0 (0.02–8.07%) and 1.87% for Ego-Exo4D (0.02–95.99%). Pooled, 21.3% of the 5,162 ground-truth boxes are small ($<$0.5% of the frame) and 52.0% are large ($>$1%), so the small-target regime in which models degrade most is a substantial minority of the data rather than a corner case.

D.3 Answer Types and Option Balance

The 2,497 answers comprise 1,192 four-way multiple-choice questions (47.7%; both planning tasks), 622 waypoint answers for both hands (24.9%; the three trajectory tasks), 404 time-stamped box-sequence answers (16.2%; object and task correspondence), and 279 open structured-JSON answers with an interaction interval and hand–object track (11.2%; HOI correspondence); the payload of each type is specified in Table 5.

Golden-option positions are balanced by the deterministic per-sample shuffle of Section C.4: in every released planning split the empirical frequencies stay within 5.4 percentage points of uniform (Table 7; Fig. 17d), and in the merged evaluated sets within 2.9 points (next-step 133/117/124/119, long-horizon 161/195/186/157 for positions A–D).

Every planning sample is additionally anchored to an exocentric-only decisive state: we verify programmatically that in all 1,192 samples the exocentric reference frame is captured exactly at the hidden-event timestamp (Section 3.3). The hidden event precedes the decision point by less than 15 s in 1,112 samples (93.3%) and by 15–60 s in the remaining 80 (6.7%). The perception and trajectory tasks are cross-view by construction—the reference view is the object of the query—so no per-sample necessity flag is defined for them.



Figure 17: Distributions of MultiView-EB beyond the headline counts (2,497 pairs; object correspondence pools ACE-Data-0+Ego-Exo4D). (a) Empirical CDFs of the target-view input duration for the five clip-based tasks (Ego-Exo4D object curve dashed); the dashed rule marks the pooled median of 30.8 s over the 1,796 clip-based samples. (b) Ground-truth box areas for object correspondence by source, with the 0.5% and 1% target-size thresholds of Section 4.4 (small/mid/large: 627/929/991 ACE-Data-0 and 472/448/1,695 Ego-Exo4D boxes). (c) Ground-truth output length per task: median (dot), interquartile range (band), and min–max (whisker). (d) Golden-option position in the merged evaluated planning sets (val+train per task) against the uniform 25% reference; neither merged set deviates by more than 2.9 percentage points (Table 7).

Table 10: Model roster: access path and per-request input budget. Local deployments serve 32,768-token context for Qwen3.5-4B, Qwen3.6-35B-A3B, and Hy-Embodied-VLM-1.0 (raised to 65,536 for the two local Qwen servers on the frame-expanded gt-aligned inputs of Section E.3 and the trajectory server), 40,960 for InternVL3.5, 131,072 for MiMo-Embodied-7B, and 262,144 for Qwen3.5-397B-A17B and the RynnBrain, RoboBrain, Cosmos3, and PhysBrain servers; gateway limits are not disclosed. Over-cap inputs are uniformly subsampled keeping first and last frames and timestamps; "no video" gateways receive clips as sampled frames; no budget entry means the full sampled input.

Model	Access	Input budget
Open-source general		
Qwen3.5-4B	local vLLM	–
Qwen3.6-35B-A3B	local vLLM	–
Qwen3.5-397B-A17B	API gateway	–
Qwen3.8-flash	API gateway	–
InternVL3.5 (4B, 8B, 38B)	local vLLM	$\leq$22 images
Proprietary general		
GPT-6-astra	API gateway	no video; $\leq$50 images
GPT-6-sol	API gateway	no video; $\leq$50 images
GPT-5.6-sol	API gateway	no video
Claude-Sonnet-5	API gateway	no video; $\leq$50 images
Gemini-3.1-pro	API gateway	–
Gemini-3.8-flash	API gateway	–
Embodied-brain models		
RynnBrain 1.1 (9B, 122B-A10B)	local vLLM	–
RoboBrain 2.5-8B	local vLLM	–
Hy-Embodied-VLM-1.0	local vLLM	$\leq$30 images
RxBrain-6B	local wrapper	$\leq$16 images; 256-tok. output
Cosmos3-Nano	local vLLM	–
PhysBrain 1.5-8B	local vLLM	–
MiMo-Embodied-7B	local vLLM	$\leq$130 images
Qwen3.5-4B-MV-Emb (ours)	local vLLM	–

D.4 Take Assignment

ACE-Data-0 ships no official split, so its 468 candidate takes are partitioned 374/94 by take with seed 42: the 374 feed the fine-tuning corpus of Appendix G, and 82 of the remaining 94 contribute at least one released pair.

The 200 Ego-Exo4D object-correspondence takes are drawn independently of the planning takes, 92 of them also appearing in the planning pool; the union of Ego-Exo4D takes across tasks is 576, so with the 82 ACE-Data-0 takes the benchmark builds on 658 takes.

E Experimental Setup

This section gives the evaluation protocol behind Section 4; verbatim instruction templates are reproduced in Appendix H, per-task answer schemas and construction details in Appendix C.

E.1 Evaluated Models

We evaluate twenty-one zero-shot models—thirteen general-purpose MLLMs (seven open-source, six proprietary) and eight embodied-brain models, listed in Table 10—plus the fine-tuned Qwen3.5-4B-MV-Emb (Appendix G). Open-weight models are served locally with vLLM—Hy-Embodied-VLM-1.0 through its official vLLM plugin, and RxBrain-6B, which vLLM does not support, behind a custom OpenAI-compatible wrapper around its official inference code; proprietary models use their inference gateways. The prompt text is identical across all models and input conditions.

E.2 Inference and Decoding

Unless a gateway forces otherwise, all tasks and models decode with one configuration: temperature 0.7, top-p 0.8, top-k 20, presence penalty 1.5, repetition penalty 1.0, thinking disabled. Generation is capped at 8,192 new tokens for correspondence and trajectory (planning 4,096 or 8,192 by deployment, immaterial since planning answers are single option letters); the local Qwen baselines fix the generation seed (3407), gateway endpoints sample server-side and cannot be seeded. Exceptions:

- **GPT-6 gateways** (GPT-6-astra, GPT-6-sol, GPT-5.6-sol): only model, messages, and maximum-token fields accepted; sampling parameters dropped, effective temperature the server default of 1.
- **Claude-Sonnet-5**: accepts the standard temperature, top-p, and top-k.
- **Gemini-3.1-pro**: thinking level HIGH, 16,384-token output budget, thoughts retained.
- **Gemini-3.8-flash**: thinking cannot be disabled; thoughts count against the 8,192-token output budget.
- **MiMo-Embodied-7B**: no thinking flag in the template; disabled via a `/no_think` suffix on the user message.
- **RynnBrain 1.1-9B**: thinking enabled on correspondence and planning (output format collapses otherwise), disabled on trajectory.
- **RxBrain-6B**: greedy decoding, 256-token output cap.
- **Trajectory tasks**: temperature 0, minimal request body (model, temperature, maximum tokens, thinking flag).

Failed or empty responses are retried up to twice and logged per sample; outputs failing to parse are re-requested up to twice where supported, then recorded as parse failures (Section E.4); deterministic GPT-gateway content-filter rejections are scored as incorrect without retry.

Media sampling is identical across models: video is decoded on uniform grids at 2 fps (planning ego clips capped at 32 frames, longer-than-16 s clips subsampled to exactly 32; correspondence clips in full, under a 2,048-frame ceiling no clip reaches), frames normalized to at most 655,360 pixels per frame (factor-32 smart resize, 16×16 patches; $1408^2 \rightarrow 800^2$). Trajectory keyframes go as native-resolution JPEGs, the grounding reference pre-sampled to eight equispaced frames. Inputs over a model's image budget (Table 10) are uniformly subsampled keeping first and last frames and timestamps; the Qwen3.8-flash gateway's 20 MB payload cap is met by higher-compression re-encoding preserving frame grid and resolution.

E.3 Input Assembly

Planning admits the viewpoint ablation of Section 4.3: *ego-only* sends the egocentric clip ending at the decision point t_d (windows up to 300 s), and *ego+exo* (Table 2) adds a single exocentric frame at t_d; the instruction is byte-identical across conditions and models. Because the shared planning template (Appendix H) mentions the exocentric frame in both conditions, any cost of the unmatched mention under ego-only depresses that baseline and can only inflate the contrast—the reported exocentric gains are upper bounds on the true benefit of the second view. **Correspondence** always presents both views: object correspondence pairs exocentric reference frames highlighting the target with a synchronized egocentric clip, HOI an exocentric interaction clip with a longer synchronized ego clip, task correspondence an exocentric multi-step reference clip with a single target-view frame. For the official object and HOI calibers the target clip is expanded into *gt-aligned* inputs (one image per ground-truth 0.5 s-grid frame, each preceded by its timestamp), so the frames seen coincide with the timestamps scored—a no-op for task correspondence's single frame and the already-timestamped Ego-Exo4D object batch. **Trajectory** grounding presents the exocentric reference clip (eight equispaced frames) with the target-view keyframe sequence; goal-conditioned prediction the target frame before the action plus exocentric frames before and after it; instruction-conditioned prediction current ego- and exo-view frames plus the instruction; waypoints lie in the image plane of each target keyframe, and every image part is preceded by its timestamp relative to the clip start. Per-task compositions and answer schemas are in Appendix C (Table 4).

E.4 Output Parsing and Scoring Conventions

Multiple choice. The option letter is extracted by a four-level cascade—(i) an explicit answer statement; (ii) a leading standalone letter; (iii) any standalone valid letter; (iv) unambiguous containment of exactly one full option text—and matched exactly against the gold letter; no match is a parse failure, counted as incorrect. **Structured outputs.** Spatial answers (time-stamped box tracks, ordered boxes with labels, hand–object tracks, waypoint groups) are parsed from JSON by increasingly tolerant extractors: whole response; each fenced code block; the balanced span at the first bracket; truncated-JSON repair; bracket-typo repair (candidates must conserve bracket counts and parse completely); any balanced span. A prose fallback extracts (box, label) pairs from numbered markdown lists for the ordered-box tasks; whatever still fails is a parse failure, with raw responses retained so scoring can be re-applied offline without re-running models. **Coordinate spaces.** Predictions may arrive in target-view pixels, in $[0, 1]$-normalized coordinates, or in the $[0, 1000)$ convention of Qwen grounding models; all three are mapped into the ground-truth pixel space (dimensions measured from the target media), and the interpretation minimizing the mean nearest-point distance to the ground truth is selected and recorded. **Worst-case scoring.** Every sample contributes to every metric (full-sample denominator): unparseable or refused outputs receive 0 on upward-bounded metrics, the worst value on downward-bounded ones (1.0 for the normalized sequence edit distance, $\sqrt{2}$—the unit-square diagonal—for all trajectory distances), and multiple-choice parse failures count as incorrect. Short refusals (e.g., "no object marked") are tracked separately from malformed outputs as a data-side signal but scored identically, so protocol breakdowns surface as worst-case scores rather than being masked.

E.5 Per-Model Protocol Notes

Reported numbers pool batches where applicable: object correspondence pools the ACE (n=125) and Ego-Exo4D (n=200) batches; every task is scored on its full released question set (n=493 next-step and 699 long-horizon planning, 279 HOI, 79 task, $77/202/343$ trajectory). All models are scored on identical manifests and prompts; the remaining per-model differences, relevant to reading Table 2, are input budgets and response-format behavior:

- **GPT-6-astra.** Frames-only video, 50-image cap.
- **GPT-6-sol.** Same gateway constraints; all eight tasks; three Ego-Exo4D content-safety rejections scored as incorrect.
- **GPT-5.6-sol.** Deterministic content-filter rejections (under 5% in affected runs) scored as incorrect without retry.
- **Claude-Sonnet-5.** No video parts (clips as sampled frames); 50-image cap (32 MB request limit).
- **InternVL3.5 (4B/8B/38B).** Correspondence subsampled to $\leq$22 images (40,960-token context); trajectory inputs ($\leq$9 images) below the cap.
- **Hy-Embodied-VLM-1.0.** $\leq$30 images (32,768-token context).
- **MiMo-Embodied-7B.** $\leq$130 images; `/no_think`; trajectory-grounding outputs collapse to a single group against eight ground-truth groups, taking the maximum DFD.
- **RynnBrain 1.1.** 1.1-9B follows the trajectory format on 71–76% of the two prediction tasks (worst-case otherwise); 1.1-122B-A10B on 94–99% but emits empty box lists on nearly all object/HOI inputs (near-zero mIoU/tIoU); 1.1-9B uses per-task thinking (Section E.2).
- **RxBrain-6B.** Free-form prose; follows the trajectory format on 5–14% (16-image cap, 256-token cap, greedy), worst-case otherwise.
- **PhysBrain 1.5-8B.** Only released conversational checkpoint of its family (PhysBrain 1.0 ships action-model weights only).

E.6 Metrics

Throughout, $\mathrm{IoU}(\cdot, \cdot)$ is the intersection-over-union of axis-aligned boxes and $\mathbb{I}[\cdot]$ the indicator function; correspondence metrics use the target view's pixel space, trajectory metrics the per-axis nor-

malized $[0,1]^2$ space (coordinates divided by the measured target-media dimensions), and sample-level quantities are averaged under the full-sample denominator. Table 4 lists each task's primary metric.

Multiple-choice accuracy (planning).

$$\mathrm{Acc} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big[\hat{\ell}_i = \ell_i^\star\big],$$

with $\hat{\ell}_i$ the parsed and $\ell_i^\star$ the gold option letter, unparseable responses counted as incorrect; the random-choice baseline is $1/C$ with $C{=}4$ options (0.25) (Qiu et al., 2026).

Object correspondence (mIoU, vIoU@τ). The predicted track is linearly interpolated to each ground-truth timestamp t_k (outside its coverage, IoU 0), following the Ego4D visual query 2D convention (Grauman et al., 2022); with g_k the ground-truth box at t_k and K_i its ground-truth frame count,

$$\mathrm{mIoU}_i \;=\; \frac{1}{K_i}\sum_{k=1}^{K_i}\mathrm{IoU}\big(\hat{b}(t_k),\, g_k\big), \qquad \mathrm{vIoU@}\tau \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big[\mathrm{mIoU}_i > \tau\big],\;\; \tau \in \{0.3, 0.5\}.$$

vIoU@τ follows the STVG convention (Zhang et al., 2020); Recovery@0.5 is the fraction of ground-truth frames with IoU ≥ 0.5.

HOI correspondence (tIoU, hand/object IoU).

$$\mathrm{tIoU} \;=\; \frac{|\hat{I} \cap I|}{|\hat{I} \cup I|},$$

the temporal IoU between predicted and ground-truth interaction intervals, following temporal-grounding retrieval (Gao et al., 2017). Each ground-truth (hand box, object box) entry is matched to the predicted entry within ± 0.5 s maximizing $\min(\mathrm{IoU}_{\mathrm{hand}}, \mathrm{IoU}_{\mathrm{obj}})$; hand (object) IoU is the mean IoU of the matched boxes. A *location hit* requires both matched boxes at IoU ≥ 0.5, a *triplet hit* additionally label agreement, following HOI-detection matching (Chao et al., 2017; Chiou et al., 2021); triplet precision counts unmatched predictions as false positives.

Task correspondence (box F1, sequence edit distance). Localization uses greedy one-to-one matching between the visible ground-truth boxes and all predicted boxes at IoU ≥ 0.5 (pairs matched in decreasing IoU, each box used at most once), following GREC (He et al., 2023), with duplicates, hallucinated objects, and boxes for invisible steps counting as false positives; with $\mathcal{M}$ the matched set, $\mathcal{G}_{\mathrm{vis}}$ the visible ground-truth boxes, and $\hat{\mathcal{B}}$ all predicted boxes,

$$\mathrm{Recall} \;=\; \frac{|\mathcal{M}|}{|\mathcal{G}_{\mathrm{vis}}|}, \qquad \mathrm{Precision} \;=\; \frac{|\mathcal{M}|}{|\hat{\mathcal{B}}|}, \qquad F_1 \;=\; \frac{2\,\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}.$$

The reported box F1 is the localization variant; the label variant additionally requires label agreement (normalized, word-order-insensitive containment with an offline-verified paraphrase cache). Order is scored by mapping each predicted label onto the first unused matching ground-truth label (unmatched labels kept distinct), giving

$$\mathrm{seqED}_i \;=\; \frac{1}{|s_i|}\,\mathrm{DL}\big(\sigma(\hat{s}_i),\, s_i\big),$$

the Damerau–Levenshtein distance normalized by the ground-truth step count (Grauman et al., 2022).

Trajectory prediction (DFD and secondary metrics). Predictions are grouped by (frame, hand label); a ground-truth group with no matching predicted group contributes the worst-case distance $\sqrt{2}$ to every distance metric, and surplus predicted groups are reported as false-positive counts.

Within a group, with predicted points $P = (p_1, \ldots, p_m)$ and ground-truth points $Q = (q_1, \ldots, q_n)$, the primary metric is the discrete Fréchet distance (DFD),

$$D_{1,j} = \max_{j' \le j} \|p_1 - q_{j'}\|, \qquad D_{i,1} = \max_{i' \le i} \|p_{i'} - q_1\|,$$
$$D_{i,j} = \max\Big(\|p_i - q_j\|,\ \min\big(D_{i-1,j},\, D_{i-1,j-1},\, D_{i,j-1}\big)\Big), \qquad d_{\mathrm{F}}(P,Q) = D_{m,n},$$

the minimum, over all order-preserving traversals of the two sequences, of the maximum point distance, robust to unequal point counts; DFD and the Hausdorff distance below follow the ShareRobot trajectory convention (Ji et al., 2025). Secondary metrics are the symmetric Hausdorff distance,

$$d_{\mathrm{H}}(P,Q) \;=\; \max\Big(\max_i \min_j \|p_i - q_j\|,\ \ \max_j \min_i \|p_i - q_j\|\Big);$$

the length-aligned RMSE, after linearly resampling the prediction to the ground-truth point count n (following VABench (Yuan et al., 2025)),

$$\mathrm{RMSE} \;=\; \Big(\frac{1}{n}\sum_{k=1}^{n} \|\hat{p}_k - q_k\|^2\Big)^{1/2};$$

the endpoint error $e_{\mathrm{end}} = \|\hat{p}_n - q_n\|$ between the final waypoints (the final-displacement-error convention (Gupta et al., 2018)); and the threshold hit rate, following TAP-Vid (Doersch et al., 2022),

$$\mathrm{hit@}\delta \;=\; \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\Big[\min_j \|\hat{p}_j - q_k\| \le \delta\Big], \qquad \delta \in \{0.02, 0.05, 0.10\},$$

whose binary caliber is $\mathrm{hit@}0.05 \ge 0.5$.

Paired significance tests. Planning condition contrasts (Section 4.3) use McNemar's exact test on paired per-sample correctness; with b samples correct only under condition A and c only under B, the exact two-sided p-value is

$$p \;=\; \min\Big(1,\ \ 2^{\,1-(b+c)} \sum_{i=0}^{\min(b,c)} \binom{b+c}{i}\Big).$$

F Per-task Error Analysis and Model Reasoning Cases

Section 4.4 classified planning failures with a 125-case manual taxonomy; this section performs the complementary analysis for the six remaining tasks: for each, we sample failed predictions from the per-sample records, assign every sampled failure an error type from a task-specific taxonomy, and illustrate the dominant types with verbatim model outputs and ground-truth overlays; two planning action-error cases from the taxonomy of Section 4.4 are shown alongside (Figure 18). Throughout, a *failure* is a per-sample event under the task's own metric—not a parse failure—and every error type describes what the model did, not how it was scored.

F.1 Labeling Protocol

Models and failure definitions. We label failures from two proprietary front-runners of Table 2—GPT-6-astra and GPT-6-sol—and one open-source model, InternVL3.5-8B, representing the small-open-model regime; trajectory-grounding failures are labeled on GPT-6-sol and InternVL3.5-8B. A sample fails when its headline metric crosses a threshold: object correspondence at per-sample mIoU < 0.3, HOI correspondence at tIoU < 0.3 *or* hand-mIoU < 0.3, task correspondence at box F1 < 0.3, and the trajectory tasks at hit@0.05 < 0.5, the scorer's per-sample criterion; unparseable outputs are protocol events, excluded from all pools (compliance rates in Section E.5).

Sampling and procedure. Small failure pools are labeled in full (all 13 GPT-6-sol ACE object failures, all 9 of its grounding failures), the rest by seeded random sample (seed 0), stratified by data source for object correspondence (ACE-Data-0 vs. Ego-Exo4D) and by failure stratum for HOI: n=50 object failures (GPT-6-sol 30: 13 ACE + 17 Ego-Exo4D; GPT-6-astra 7; InternVL3.5-8B

Next-Step Planning Action Errors

(a) Next-Step Planning: Progress Behind - gpt-6-astra (progress behind)

Ego history:



t = 53 s t = 120 s t = 192 s t = 262 s

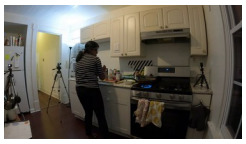
Goal:

Add oil to skillet and adjust heat

Question:

What should the camera wearer do next to achieve the goal?

Exo evidence:



exo · now

Options:

A. adjust knob of gas cooker (on the counter to the left)

B. place oil bottle on countertop (on the countertop next to the sink)

C. open lid of oil bottle (on the counter to the right of the sink)

D. adjust knob of gas cooker (on the front panel of the stove)

Error: Re-puts the oil bottle on the counter (done 0.8s before the head turn) instead of adjusting the gas knob.

(b) Next-Step Planning: Progress Ahead - Gemini-3.1-pro (progress ahead)

Ego history:



t = 6 s t = 14 s t = 23 s t = 31 s

Goal:

Juggle and exchange passes with partner X and O

Question:

What should the camera wearer do next to achieve the goal?

Exo evidence:



exo · now

Options:

A. receive ball (in the hands of the person standing to the right)

B. juggle ball (on the green turf in front of the yellow goal)

C. pass ball (on the green turf near the yellow goal)

D. receive ball (in the hands of the person in the red shirt)

Error: Skips receiving the partner's pass and jumps to passing the ball.

Gemini-3.1-pro reasoning (verbatim):

[…] This directly corresponds to the action of "pass ball". Considering the available options: A. receive ball B. juggle ball C. pass ball D. receive ball The action that is visually confirmed and ongoing at the point of observation (00:28-00:31) is the act of passing the ball. Therefore, option C, "pass ball," accurately describes the current and concluding action from the observed moments, which is precisely what the prompt is asking for within the context of the "current observation" leading into the next possible state. The prompt is not asking for the *subsequent* action to the pass, but rather the action that is currently unfolding or has just been completed.

Figure 18: Next-step planning action-error cases from the 125-case taxonomy of Section 4.4, in the case layout of Figure 19 (options verbatim; green fill: gold option; pink fill with a red marker: the model's chosen option; gray: remaining distractors; the reasoning excerpt is verbatim from the archived output). **(a)** Progress behind, gpt-6-astra, `georgiatech_cooking_02_02_2_k23`: re-puts the oil bottle on the counter (done 0.8 s before the head turn) instead of adjusting the gas knob (gold D, chosen B). **(b)** Progress ahead, Gemini-3.1-pro, `iiith_soccer_002_4_k3`: skips receiving the partner's pass and jumps to passing the ball (gold A, chosen C); the archived reasoning shows the model anchoring on the action currently unfolding at the decision point rather than the next one. Error gists are quoted verbatim from the manual taxonomy.

Table 11: Object-correspondence error taxonomy over 50 sampled failures (per-sample mIoU < 0.3; parse failures excluded). "astra+sol" pools GPT-6-astra (n=7) and GPT-6-sol (n=30, ACE + Ego-Exo4D); counts are failures, percentages of the column; one type per failure, by a documented priority rule (size and occlusion signatures first, then binding level).

Error type	Definition	astra+sol n=37	InternVL n=13	All n=50
Misbinding	Tracks a different object or region: best IoU < 0.05 with the target and boxes sit $\geq$8% of the frame away (render-verified)	13 (35%)	2 (15%)	15 (30%)
Near miss	Boxes touch the target (best IoU 0.05–0.3) or sit within 8% of the frame of it, but never lock on (best IoU < 0.3)	14 (38%)	2 (15%)	16 (32%)
Intermittent binding	Object bound (IoU ≥ 0.3) in part of the clip, lost elsewhere	5 (14%)	4 (31%)	9 (18%)
Phantom boxes	Boxes emitted while the object is out of view (GT track has an occlusion gap)	3 (8%)	0	3 (6%)
Static box	One location repeated at every timestamp while the target moves	0	5 (38%)	5 (10%)
Size miscalibration	Bound and centered but systematically mis-scaled ($\geq 2\times$ or $\leq 0.5\times$ GT area)	2 (5%)	0	2 (4%)

13), n=50 HOI (27/13/10), n=50 task correspondence (22/15/13), and n=89 trajectory (grounding 25, goal-conditioned 32, instruction-conditioned 32). Per-case features are recomputed from the archived predictions with the official scorer's own parsers, label predicate, and coordinate-space convention (scripts and per-case evidence in `figures/appendix/error_cases/`). Each failure receives one primary error type under a documented priority rule (task correspondence is multi-label, its types co-occurring almost universally), assigned from rule-based pre-labels; categories that cannot be separated numerically—which object a misbound track follows, whether far-away boxes sit on a real object or on empty scene regions, whether a residual trajectory error is a wrong shape—are adjudicated visually on rendered overlays of the model's actual input frames (green ground truth, red prediction). All 45 such object cases, 23 HOI cases, and 18 residual trajectory cases were adjudicated this way.

F.2 Object Correspondence

Object correspondence fails at binding, not tracking (Section 4.4): 62% of the pooled failures are spatial mislocalizations of roughly equal split—misbinding onto a coherent wrong object (Figure 19(a)) or near misses that hover near the target without ever locking on—concentrated in the head models (GPT-6-astra fails 7 of 125 ACE samples, GPT-6-sol 13 of 125 ACE and 64 of 200 Ego-Exo4D). InternVL3.5-8B instead repeats a single location at every timestamp or binds only intermittently, consistent with the multi-frame binding collapse of Section 4.4; on Ego-Exo4D its boxes often sit on empty field or sky. Temporal and calibration errors are rare (three occlusion-gap phantom tracks, two scale errors); Figure 19(b) shows both at once: GPT-6-astra binds the cup while visible, emits boxes through the 9 s occlusion the instruction asks it to omit, and never re-binds (mIoU 0.27).

F.3 HOI Correspondence

Models find *when* an interaction happens more reliably than *where* (tIoU up to 0.83 vs. hand/object IoU at most 0.55/0.58); from the failure side (Table 12), among the head models' sampled failures the window is the more common culprit (19 of 40)—window errors are rare in absolute terms (GPT-6-sol: 15 of 279 samples) but dominant within its failure pool. The window errors are directional: 16 of 24 place the interaction too early (median center shift -0.8 s), 23 of 24 are shorter than the ground truth (median 0.70×), and 8 begin exactly at the clip start—the model anchors on the first plausible interaction and truncates it. The spatial residual splits into imprecise boxes near the correct hand–object pair (12), render-verified wrong objects or regions (7), and hand-identity swaps (4); object *naming* is not a meaningful failure axis, the flagged "mismatches" being paraphrases such as "yellow mug" for the ground-truth "Cup Yellow" while the boxes are the problem.

Figure 19(c) and the transcript below show a typical case: GPT-6-sol tracks a "mug" from the first second of the clip while the referenced interaction occupies $[2.6, 9.6]$ s.



Figure 19: Correspondence failure cases, one case per block (left: the model's actual input frames, ground truth in green and prediction in red as vector overlays; right: question, ground-truth summary on green, verbatim prediction excerpt on pink, error gist). **(a)** Object, GPT-6-sol, take-000262 (mIoU 0.00): tracks a different object than the spatula marked in the exocentric reference. **(b)** Object, GPT-6-astra, take-000179 (mIoU 0.27): bound while visible, phantom boxes through the 9 s occlusion, no re-binding after reappearance. **(c)** HOI, GPT-6-sol, take-000307 (tIoU 0.09): predicted window $[0.0, 3.5]$ s vs. ground truth $[2.6, 9.6]$ s. **(d)** Task, GPT-6-sol, take-000163 (box F1 0.22): three ground-truth interacted objects vs. seven narrative steps enumerated from the instruction.

Table 12: HOI-correspondence error taxonomy over 50 sampled failures (per-sample tIoU < 0.3 or hand-mIoU < 0.3; parse failures excluded), stratified sampling with all temporal failures of GPT-6-sol and GPT-6-astra included; counts are failures, final column percent of all 50. Wrong-window samples fail tIoU; all others localize the interval but fail the hand–object track.

Error type	Definition	GPT-6-sol n=27	GPT-6-astra n=13	InternVL n=10	All n=50
Wrong window	Interaction interval misplaced (tIoU < 0.3)	15	4	5	24 (48%)
Imprecise boxes	Interval correct; hand/object boxes near the GT pair but IoU < 0.3 (median center offset 12% of the frame)	5	5	2	12 (24%)
Wrong instance	Interval correct; boxes on a different object/region (>15% of the frame away; render-verified)	1	4	2	7 (14%)
Hand identity	Named hand opposite to the GT hand (tracked geometry belongs to the other hand)	3	0	1	4 (8%)
Label variant	Object named with a paraphrase (e.g. "yellow mug" for "Cup Yellow") while spatially off	3	0	0	3 (6%)

Table 13: Task-correspondence error types over 50 sampled failures (per-sample box F1 < 0.3; parse failures excluded); multi-label, types co-occurring in 49 of 50 failures, so counts sum to more than n. "Named" uses the benchmark's containment label predicate, so a step missed by name may still be covered by a paraphrase.

Error type	Definition	GPT-6-sol n=22	GPT-6-astra n=15	InternVL n=13	All n=50
Missed steps	$\geq$1 GT step never named by any prediction (containment predicate)	20	15	13	48 (96%)
Extra steps	$\geq$1 predicted step matches no GT step	22	15	13	50 (100%)
Localization off	A name-matched step exists and the mean box IoU over name-matched steps is < 0.3	7	8	2	17 (34%)
Null on visible	A named, visible GT step answered with `null`	0	2	4	6 (12%)
Box on invisible	A box emitted for a named, invisible GT step	2	1	0	3 (6%)

```
GPT-6-sol, take-000307 (tIoU 0.09; GT window [2.6, 9.6]):
{"time_range":[0.0,3.5],"hoi_track":[{"time":0.1,
"object_bbox_2d":[392,403,442,463],"object_label":"mug",
"hand_bbox_2d":[465,491,570,635],"hand_label":"right hand"},
{"time":0.6, ... , "object_label":"mug", "hand_label":"right hand"},
...]}
```

Because the evaluated models emit answer JSON without visible deliberation, we also quote RynnBrain 1.1-9B, the one evaluated model whose full chain of thought is archived alongside its predictions—the same early-anchoring pattern with the mechanism visible: a narrated confident match to the first event in the clip and a 0.5 s window for a 5 s interaction (tIoU 0.0, hand-mIoU 0.02).

```
RynnBrain 1.1-9B, take-000074 (GT window [1.05, 6.05]; output [0.45,
0.92]):
[...]  2.  Analyze the Egocentric Clip:  ...Around 0.4s to 0.9s,
the person is seen pouring water from the cup into the sink.  [...]
5.  Refine the Time Range:  ...the pouring starts around 0.45s
and finishes by 0.92s.  This aligns well with the duration of the
exocentric action.
```

F.4 Task Correspondence

Task-correspondence failure is step-retrieval before geometry (Table 13): every sampled failure misses ground-truth steps and adds spurious ones (96%/100%), the median failed prediction names only 15% of the reference clip's interacted objects, and 35 of 50 score box F1 exactly 0. Naming source is not the discriminator—successful and failed predictions alike overwhelmingly label objects mentioned in the task instruction (92% vs. 83% of predicted labels)—but the ground truth is the specific object set actually manipulated in the reference clip, and failed predictions enumerate the

Table 14: Trajectory error taxonomy over 89 sampled failures (per-sample hit@0.05 < 0.5; parse failures excluded), by task; one type per failure, by priority (hand identity, omitted group, static, truncation, displacement, residual wrong shape). Grounding emits per-keyframe hand groups (an omitted group is one frame's hand track), the prediction tasks one path per hand (an omitted group is an entire hand). "Displacement" requires shape and scale preserved (spread within 30% of GT, path length $\geq 0.6\times$) but the path translated by $\geq 10\%$ of the image.

Error type	Definition	Grd. n=25	Goal n=32	Instr. n=32	All n=89
Omitted group	A hand/frame group absent entirely (scored at the maximum distance)	9	6	4	19 (21%)
Truncated path	Moving hand's path far shorter than GT (length $< 0.4\times$)	4	10	4	18 (20%)
Wrong shape	Prediction moves at comparable scale but along a different path (render-verified)	0	9	9	18 (20%)
Displacement	Shape and scale preserved, path translated $\geq 10\%$ of the image	0	5	12	17 (19%)
Hand identity	Path geometry correct but attributed to the opposite hand (min. point distance < 0.10 to the opposite GT path)	10	2	3	15 (17%)
Static collapse	Near-static points where the GT hand moves (spread $< 1\%$ of the frame)	2	0	0	2 (2%)

instruction's narrative stations instead. Visibility calls are rarely decisive (12% `null` on a visible named step, 6% box on an invisible one), and order is never the sole defect—all 12 failures retrieving at least two correctly named steps match the ground-truth order—mirroring Section 4.4: models read *what kind* of activity is underway from language, but not *which specific objects* the reference view shows.

```
GPT-6-sol, take-000163 (GT steps:  Tea Box1 / Coffee Cup / Water
Dispenser):
[{"step":1,"bbox_2d":[221,173,260,205],"label":"foil wrapping
on the kitchen island"}, {"step":2,...,"label":"mug"},
[...]  {"step":5,"bbox_2d":[694,309,800,577],"label":"sofa"},
{"step":6,"bbox_2d":null,"label":"TV"}, {"step":7,...,"label":"sink
and faucet for washing the mug"}]
```

F.5 Trajectory Prediction

The three trajectory tasks fail differently, tracking how much of the answer the input determines (Table 14). On grounding—where the reference clip shows the motion to copy—failures are structural: InternVL3.5-8B accounts for 9 of the 10 hand-identity confusions and 5 of the 9 omitted groups, its confusions pure identity swaps (Figure 20(a): correct right-hand geometry labeled `left_hand`, right-hand groups omitted); GPT-6-sol, the strongest grounding model in the labeled set (DFD 0.146, behind only GPT-6-astra and GPT-5.6-sol overall), splits its nine grounding failures between omitted groups (four) and truncated paths (four), truncation also dominating its pooled trajectory failures (13 of 33). On the two prediction tasks—where the motion must be anticipated, not copied—errors shift toward geometry: goal-conditioned failures split between truncated paths (10) and wrong shapes (9), and instruction-conditioned failures are dominated by displacement (12), a right-shaped path translated to the wrong region; Figure 20(b) shows the truncation mode. Across all 89 failures the distribution is notably flat—no single type exceeds 21%—once a model produces a parseable trajectory, its errors spread almost evenly across omission, truncation, distortion, displacement, and identity; the two model-level clusters are the open model's identity/omission pattern on grounding (14 of 16 failures) and the strongest model's motion under-extension on prediction (13 of 33).

G Fine-tuning on MV-EmData

This section details the fine-tuning row of Table 2: the construction and composition of the MV-EmData corpus, the training recipe, and full before/after results on all eight tasks. We fine-tune Qwen3.5-4B (Qwen Team, 2026a) on MV-EmData and refer to the result as Qwen3.5-4B-MV-Emb (Qwen3.5-4B fine-tuned on MV-EmData). The model is evaluated with the baseline Qwen3.5-

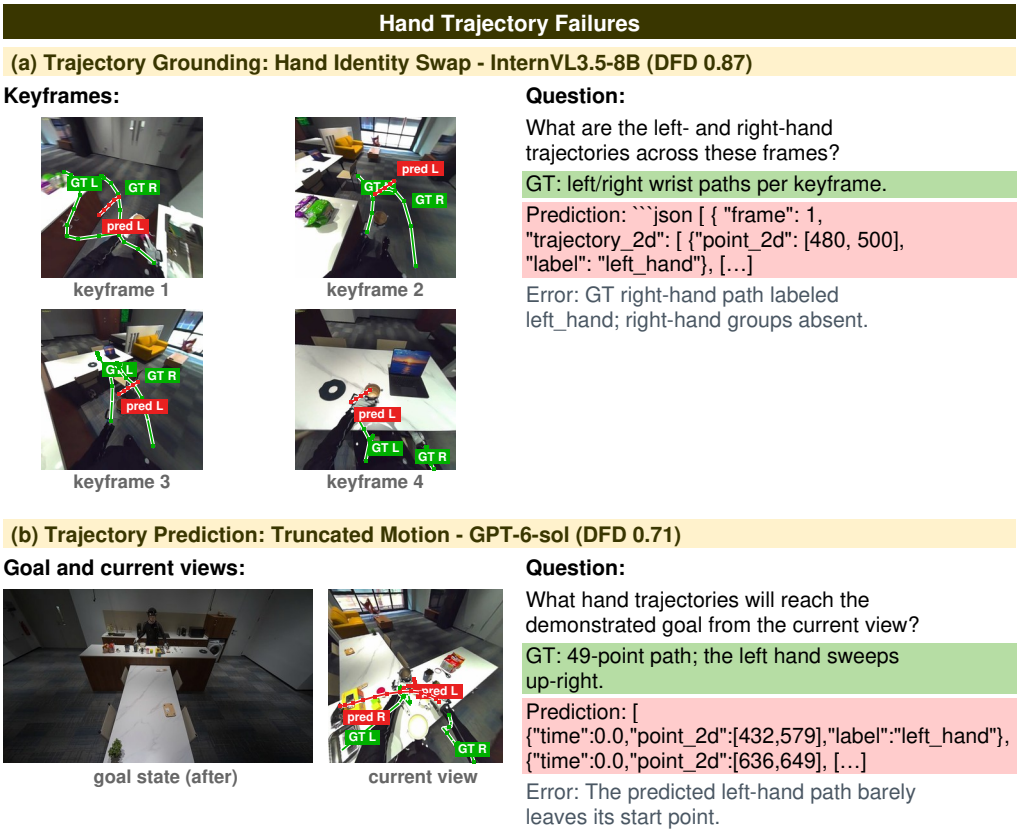


Figure 20: Trajectory failure cases, in the case layout of Figure 19 (green: ground truth; red: prediction). **(a)** Grounding, InternVL3.5-8B, take-000074 (DFD 0.87): follows the GT right-hand path in every keyframe but labels it `left_hand`, and the right-hand groups are absent—correct geometry, wrong hand identity. **(b)** Goal-conditioned prediction, GPT-6-sol, take-000434 (DFD 0.71): given the goal state, the predicted left-hand path barely leaves its start point where the ground truth sweeps up and to the right.

4B protocol verbatim—same manifests, decoding, parsing, and scoring (Section E)—so every before/after contrast below isolates the effect of the fine-tuning data.

G.1 Corpus Construction and Composition

MV-EmData provides supervised data for the eight tasks of MultiView-EB: question pools with the benchmark's input structure and answer schemas, drawing on all three ground-truth tracks of Table 6—mocap projection for the six ACE-Data-0 tasks, native Ego-Exo4D annotations for object correspondence, planning, and one trajectory track, and the self-built detection–tracking–naming stack for Ego-Exo4D HOI, task correspondence, and a second, much larger trajectory track. The self-built pools extend coverage to task–source combinations the benchmark does not evaluate: HOI and task correspondence are ACE-only at evaluation, as are all three trajectory tasks. Sixty-one ACE samples flagged by the QR synchronization audit (Section C.6) are excluded at conversion.

Each sample becomes one supervised chat turn: the user message lists the input views in order, each frame preceded by its timestamp in seconds (the same gt-aligned rendering as the evaluation inputs, Section E.3), followed by the task instruction, and the assistant message is the ground-truth answer serialized as compact JSON, with multiple-choice answers storing the golden option text rather than its letter and spatial coordinates rescaled from the target view's pixel space into the model's dimension-relative $[0, 1000)$ representation, the convention of the Qwen grounding family.

Table 15: Composition of the MV-EmData training mix: supervised samples per task and ground-truth track, after conversion, QR-taint exclusion, the per-dataset 1% validation holdout, and the 16,384-token length filter (29 over-length HOI rows dropped). The three E4D self-built trajectory sets are downsampled to 3,000 converted rows each (2,970 shown) from 15,250, 16,492, and 11,050; without the downsampling the conversion totals 72,195 rows. "–" marks combinations with no samples.

Task	ACE-Data-0	E4D native	E4D self-built	Total
Object corr.	445	4,084	–	4,529
HOI corr.	901	–	5,950	6,851
Task corr.	281	–	2,931	3,212
Next-step	–	8,852	–	8,852
Long-horizon	–	2,438	–	2,438
Traj. grounding	278	395	2,970	3,643
Traj. pred. (goal)	576	431	2,970	3,977
Traj. pred. (instr.)	1,212	313	2,970	4,495
Total	**3,693**	**16,513**	**17,791**	**37,997**

The conversion covers 17 per-task–source datasets and produces 72,195 rows at full size, of which the three self-built trajectory sets account for 42,792; each of the three is downsampled to 3,000 converted rows to balance the task mixture, and a 1% per-dataset validation holdout plus the 16,384-token sequence filter (29 over-length HOI rows) leave **37,997 training rows** (Table 15).

G.2 Training Configuration

All runs share one recipe, implemented with the ms-swift trainer on the Megatron backend: all parameters of Qwen3.5-4B, vision tower and aligner included, are fully fine-tuned with AdamW, weight decay 0.1, gradient clipping 1.0, and a learning rate of 10^{-5} cosine-decayed to 10^{-6} after 5% warmup. The global batch is 64 sequences (per-device micro-batch 1, 32-way data parallelism, two accumulation steps) packed to 16,384 tokens with bin-packing, the loss computed on assistant tokens only; one epoch over the 37,997 rows is 12,980 packed sequences of mean length 14,773 tokens and 202 optimizer steps, and the final checkpoint of the epoch is used. A run takes about 20 minutes on 32 GPUs (four nodes).

G.3 Results

Table 16 reports all eight tasks before and after fine-tuning, under the ego+exo condition and the same calibers as Table 2; Table 17 reports two retraining variants under the same protocol.

Planning. Accuracy rises by 24.9 and 20.4 points to 0.545 and 0.602, approaching the strongest zero-shot models (best zero-shot: 0.582 next-step and 0.720 long-horizon, GPT-6-astra; 0.639 long-horizon, GPT-6-sol). The gains hold across question pools (fine-tuned accuracy 0.540–0.557 next-step and 0.598–0.611 long-horizon), and the planning portion of the fine-tuning corpus shares no take with the evaluation questions, so this is cross-take generalization, not memorization.

Correspondence. All three tasks improve by large margins, on both sources for object correspondence (ACE 0.067→0.542 mIoU, Ego-Exo4D 0.107→0.487). The baseline leaves 51%/14%/8%/46% of its object-ACE/object-E4D/HOI/task responses unparseable—worst-case scored, Section E.4—whereas the fine-tuned model parses 100% of all correspondence and trajectory outputs; the trajectory tasks, where the baseline already parsed 99–100%, are the clean geometry-only read, and planning parse rates dip only from 1.00 to 0.96–1.00. Task correspondence shows over-generation—7.7 boxes per sample against 3.5 visible ground-truth steps (baseline 1.5 against 2.7)—so the sequence edit distance worsens from 1.06 to 1.11 even as box F1 rises from 0.012 to 0.443. After fine-tuning, the 4B model would rank third on object correspondence (behind GPT-6-astra 0.660 and GPT-6-sol 0.556) and task correspondence (0.539/0.478), second on HOI tIoU (0.825), with HOI triplet F1 rising from 0.004 to 0.428.

Table 16: Qwen3.5-4B before and after fine-tuning on MV-EmData, ego+exo condition, calibers identical to Table 2 (object correspondence pools the ACE and Ego-Exo4D batches; trajectory DFD is the final rescored caliber). Δ = fine-tuned $-$ baseline; for the $\downarrow$ metrics (sequence edit distance, DFD) negative Δ is an improvement.

Task (n)	**Metric**	**Qwen3.5-4B**	**MV-Emb**	Δ
Object corr. (325)	mIoU	0.092	0.508	$+0.416$
	vIoU@0.5	0.037	0.594	$+0.557$
HOI corr. (279)	tIoU	0.491	0.740	$+0.249$
	hand IoU	0.093	0.458	$+0.365$
	object IoU	0.079	0.496	$+0.417$
Task corr. (79)	box F1	0.012	0.443	$+0.431$
	seq. edit dist. ($\downarrow$)	1.06	1.11	$+0.05$
Next-step (493)	accuracy	0.296	0.545	$+24.9$ pt
Long-horizon (699)	accuracy	0.398	0.602	$+20.4$ pt
Traj. grounding (77)	DFD ($\downarrow$)	0.643	0.512	-0.131
Traj. pred. goal (202)	DFD ($\downarrow$)	0.623	0.361	-0.262
Traj. pred. instr. (343)	DFD ($\downarrow$)	0.737	0.412	-0.325

Table 17: Retraining variants with modified trajectory conversion. All three models are full one-epoch retrains over the same 17 datasets (Section G.2); only the converted trajectory data differs. Trajectory metrics are DFD ($\downarrow$), final rescored caliber, n=77/202/343.

Variant (training change)	**Grd.**	**Goal**	**Instr.**
Base (reported; builder prompts, full-frame grounding inputs)	0.512	0.361	0.412
instrfix (evaluation-verbatim trajectory instructions)	0.522	0.384	0.391
gndalign (additionally aligned grounding presentation)	0.528	0.364	0.426

Trajectory prediction. All three tasks improve: grounding by 0.13 DFD, goal- and instruction-conditioned prediction by 0.26 and 0.33, with the secondary metrics moving in step (goal hit@0.05 0.13$\rightarrow$0.35, instruction endpoint error 0.71$\rightarrow$0.38). The fine-tuned model's goal and instruction DFDs (0.361/0.412) are bettered only by the two strongest proprietary models (GPT-6-astra 0.306/0.373, GPT-5.6-sol 0.356/0.393), with Hy-Embodied-VLM-1.0 additionally ahead on goal-conditioned prediction (0.352).

H Complete Prompt Packages

This section reproduces the prompts used to construct and evaluate MultiView-EB, quoted verbatim from the pipeline source and the released manifests; Appendix C locates each stage in the pipeline and Appendix E gives the evaluation protocol. Interpolated builder expressions appear as a {slot}; all other text is byte-identical. Bracketed ellipses ([...]) replace only enumerated per-sample content. Prompts are single paragraphs in the source; line breaks inside a box are typesetting artifacts.

Construction uses GPT-5.6 for the goal-coherence curation pass, Gemini 3.8 Flash for the remaining text-side judges (goal-tree parsing, clustering, relevance, temporal gating, decoy generation), and qwen3.8-flash for vision-side calls (grounding, visibility, naming, refinement); the blind decoy panel mixes two Qwen text models with one Gemini seat, so no family certifies its own decoys, and GPT-6-astra appears once, in ground-truth refinement (Sec. H.1.6).

Table 18: Construction prompts not quoted verbatim in this appendix. Unless noted, all calls run at temperature 0; "3-judge panel" = qwen3.8-flash, qwen3.5-plus, and Gemini 3.8 Flash with majority voting. Script names give the provenance of each prompt.

Prompt	Stage	Model	Description
Purposeful-goal filter	P1 curation	GPT-5.6	Judges each goal window purposeful and keeps its real serving steps; only purposeful goals with 4–20 kept actions become question goals (`goal_tree_v2.py`)
Atomic-action rewrite	P1 curation	Gemini 3.8 Flash	Rewrites kept narrations as unified lowercase verb–object phrases that become the option surface strings; raw timestamps kept as the time skeleton (`goal_tree_v2.py`)
Semantic-cluster sampling	P3 distractors	Gemini 3.8 Flash	Clusters same-goal candidate windows semantically; one distractor sequence per non-golden cluster, no look-alike fallback (`factorial_options.py`)
Repeatable-action gate	P3 distractors	Gemini 3.8 Flash	Keeps the already-completed-action distractor only if doing it again now would still be reasonable (`temporal_distractors.py`)
Location grounding, golden	P3 options	qwen3.8-flash	One true and one baseline-wrong parenthetical location clause per golden step, from ego frames at $t-2$, $t-1$, t and an exo keyframe (`factorial_options.py`)
Location grounding, distractors	P3 options	qwen3.8-flash	True-only location clauses for distractor sequences over the same frame layout (`factorial_options.py`)
Duplicate-instance hint	P3 options	Gemini 3.8 Flash	Appended to decoy generation when a probe flags near-identical target instances; steers decoys to the other instance (`adversarial_locations.py`)
Necessity gate and probes	P3 gating	qwen3.8-flash	Strict-ego and generous-exo visibility of the deciding location (kept only if ego-invisible and exo-visible), with one-line ego-detect and exo-salience probes (`det_gate.py`)
Step–object matching	Task-corr. build	qwen3.8-flash	Assigns each step its target and tool objects from the VLM-named track labels; a companion vision call grounds them on the target frame (`build_qa_fast.py`)
Trajectory–instruction merge	Instr. build	vLLM Qwen3-VL	Rewrites a clip's chronological atomic descriptions into one second-person imperative instruction (Ego-Exo4D; the released ACE batch takes instruction text from the dataset captions) (`build_qa.py`)
Privileged-info. judge	QA filtering	3-judge panel	Re-examines each distractor given ground-truth facts (four fact lines incl. negative and positive controls; also run without the fact line) (`judge_ambiguity.py`)
Golden-target audit	QA filtering	3-judge panel	Judges the designated answer itself—plan-step worth, coherence, predictability, artifact—with a keep/drop verdict (`judge_ambiguity.py`)
Constructive uniqueness	QA filtering	gen. + 2 blind verifiers	Model A proposes one concrete alternative next action (or NONE); two blind models verify it; a doubly verified alternative removes the sample (`verify_constructive.py`)
Occlusion-type audit	QA filtering	3-judge panel	Re-judges kept samples over an invisibility taxonomy; a majority `actually_visible` verdict flags a gate false positive (`audit_occlusion.py`)
Sweep stage 1: blind description	Instr. sweep	Gemini 3.8 Flash	Describes eleven evenly spaced frames per view with burned-in timestamps; the neutral descriptions feed the stage-2 judge quoted below
Per-track box refinement	GT refinement	qwen3.8-flash	Refines mocap-projected boxes on zoomed crops in $[0, 1000)$ coordinates under five rules; replies gated against the original box (`ace_bbox_refine.py`)
Exo reference-box refinement	GT refinement	GPT-6-astra	Re-localizes the exocentric reference marking box from the clean full-resolution frame given only the object label (`ace_exo_refine_gpt6.py`)

H.1 Benchmark-Construction Prompts

The construction prompts behind the paper's central claims are quoted verbatim below; every other construction prompt is one row of Table 18, and the full texts of all prompts are included in the code release.

H.1.1 Phase 1: goal extraction and curation

Hierarchical goal extraction. Per take, the timestamped narrations are segmented into a three-level goal tree (overall → clip → sub-clip); the user message is the narration table, with intervals extended from point events per Exo2Ego (`phase1_segmentation/goal_hierarchy.py`, `SYSTEM_PROMPT` and `build_user_prompt`; temperature 0).

```
### The following provides timestamped narrations about actions from an ego-centric
video.
Your tasks are:

1. Carefully read the provided narrations of this ego-centric video, and connect the
narrations by timestamps. While reading the narrations, please pay attention to
```

```
inferring the temporal relationships of actions, pre- and post- conditions of actions,
object affordance, object states, and possible changes of functional zones that
support different activities, etc.

2. Recognize the overall goal of the actions in this ego-centric video.

3. Temporally segment this video into clips. Actions in the same clip complete the
same subgoal that is helpful for achieving the overall goal.

4. Further segment each clip into more fine-grained sub-clips. Actions in the same
sub-clip are likely to take place in a compact space of the same functional zone
without moving around, and focus on a secondary subgoal that is helpful for
achieving the corresponding parent-level subgoal.

### All the related timestamps in the output should be copied from the given
timestamped narrations; the descriptions of the overall goal, subgoals, and secondary
subgoals should be detailed with related essential objects; and the output must
follow the JSON format:

{"overall_goal": ... , "clips": [{"subgoal": ..., "start_timestamp": ...,
"stop_timestamp": ..., "sub_clips": [{"secondary_subgoal": ..., "start_timestamp":
..., "stop_timestamp": ...}, ...]}, ...]}

Respond with the JSON object only, no other text.

-----
Begin!

### Timestamped narrations are:

start_timestamp               |          stop_timestamp                 |
narration
[... one row per narration interval, in the same fixed-width three-column layout ...]
```

H.1.2 PHASE 2: SELF-BUILT ANNOTATION VLM CALLS

Object-track naming and hand-side classification. Each BYTE-associated object track is named from its top representative crops (majority vote over up to three); hand fragments are classified left/right (majority over up to two) and merged into one track per side, with a take-level consistency check (`phase2_base_annotation/selfbuilt/hoi_fasttrack.py`; qwen3.8-flash, temperature 0, thinking disabled).

Naming (system message):

```
You name objects in cropped video frames for object tracking. Return only valid JSON:
{"object_name": "short lowercase noun phrase"}.
```

Naming (user message, one per crop):

```
Name the salient object being handled in the image.
```

Hand-side classification (user message, one per crop):

```
Is this the camera wearer's LEFT or RIGHT hand? Answer ONLY JSON: {"side": "left"} or
{"side": "right"}. If unclear, guess.
```

H.1.3 PHASE 3: PLANNING QA CONSTRUCTION

Hidden-event relevance judge. At every mined decision point, the ego-invisible exo event anchoring the sample must be judged necessary for predicting the future step(s); failures are dropped (`tasks/planning_next_step_v3/build_qa.py`, `relevance_judge`; Gemini 3.8 Flash, temperature 0).

```
You are strictly validating a multi-view egocentric PLANNING benchmark. The event
below was observed ONLY from a third-person (exocentric) camera and is invisible in
the wearer's first-person view. The QA asks a model to PREDICT the FUTURE step(s)
listed below.

Hidden (exo-only) event: "{hidden event text}"
(context) current step being finished: "{current step}"
FUTURE step(s) to predict: {future steps}

Judge relevant=true ONLY IF knowing this hidden event is genuinely NECESSARY or
materially helpful to correctly predict/choose the FUTURE step(s) above -- i.e.
there is a concrete causal or object/state link to those future actions. If it only
describes the current step, or is generic co-occurrence, or you are unsure, answer
relevant=false.
Return ONLY JSON: {"relevant": true|false, "reason": "<=15 words"}
```

Adversarial decoy generation and blind gate. For every golden step, decoy locations that are real, equally plausible a priori, and genuinely different are proposed by Gemini 3.8 Flash from the same grounding frames (temperature 0.4); each decoy is paired with the true location at randomized sides and shown, without any image, to a three-seat text-only panel (qwen3.8-flash, qwen3.5-plus, Gemini 3.8 Flash, temperature 0), and the decoy whose panel accuracy is closest to chance is kept (`adversarial_locations.py`, `GEN_PROMPT`/`BLIND_PROMPT`).

```
These are frames from an egocentric task scene. The first {n_ego} image(s) are the
wearer's FIRST-PERSON frames leading up to the moment of interest, in time order.
The final image is the THIRD-PERSON (exocentric) view of the same scene at that
moment.

The person's action is: "{action}"
The target's TRUE location in this scene is: "{true_loc}"

Propose {k} DECOY locations to be used as WRONG answers in a benchmark. The
benchmark tests whether a model can SEE where the object is -- so the decoys must be
impossible to rule out without looking at the scene.

Each decoy must satisfy ALL of:
  1. REAL -- the place actually exists in this scene (you can see it in the images).
  2. EQUALLY PLAUSIBLE A PRIORI -- for this kind of object in this kind of room, it
     is just as natural a resting place as the true location. Someone who is told
     ONLY the action and the two candidate locations, WITHOUT seeing the room, must
     NOT be able to tell which is correct.
  3. GENUINELY DIFFERENT -- a different surface/furniture from the true location,
     not the same place reworded.

Avoid decoys that common sense rejects (e.g. a cooking tool "on the dining table"
when the true location is a kitchen rack -- everyone knows tools live in the
kitchen). Prefer decoys drawn from the SAME functional area as the true location: a
different counter, the other shelf of the same rack, the adjacent drawer, a nearby
hook.
{dup_hint}
Write each as a short prepositional phrase in words, no coordinates, no numbers.

Return ONLY JSON: {"decoys": ["...", "...", ...]} with exactly {k} distinct entries.
```

Blind A/B decoy gate (`BLIND_PROMPT`):

```
A person is about to perform this action: "{action}"

Where is the target of that action located? Choose the single best option:
A. {a}
B. {b}

You have no picture of the scene -- answer from general knowledge about where such
things are usually kept. Answer with the single option letter only.
```

The remaining planning-stage prompts (semantic-cluster sampling, repeatable-action gate, location grounding, duplicate-instance hint, necessity gate and probes) are registered in Table 18.

H.1.4 Phase 3: perception and trajectory QA construction

The task-correspondence builder (step–object matching plus on-frame grounding) and the Ego-Exo4D trajectory-instruction merge are registered in Table 18.

H.1.5 QA filtering and auditing

These prompts implement the consistency-verification and vision-grounded-filtering gates of Sec. C.6. The other gates—the privileged-information distractor judge, the golden-target audit, the constructive uniqueness check, and the occlusion-type audit—are registered in Table 18; unless noted, each runs at temperature 0 with three independent judges (qwen3.8-flash, qwen3.5-plus, Gemini 3.8 Flash) and majority voting.

Visual-verifiability filter. The vision-grounded filter inspects the probe frames of every planning sample and keeps only items whose target object itself is visible at the deciding location (`filter_visual_verifiable.py`; input = the five ego probe frames plus the exocentric observation frame).

```
IMAGES 1-{n} are the camera wearer's FIRST-PERSON frames, in time order.
IMAGE {x} is the THIRD-PERSON view of the same scene at the current moment.

The wearer's next action is: "{action}"
```

```
A benchmark claims the target of that action is located: "{loc}"

A model answering this benchmark must be able to VERIFY that claim from the
pictures. Decide which of these is true, judging mainly from IMAGE {x} (the
third-person view), using the first-person frames only as extra context:

  object_visible        -- the TARGET OBJECT ITSELF is visible at that location;
                           you can point at it in the picture.
  container_opaque      -- the location is a container or enclosure you CAN see
                           (fridge, closed cabinet, drawer, opaque bag/basket,
                           covered box), but its CONTENTS are not visible, so you
                           cannot confirm the object is inside.
  location_only         -- the location/surface is visible, but the target object
                           is not discernible there (too small, blocked, out of
                           frame, or simply absent from view).
  location_not_visible  -- the stated location itself cannot be seen in these
                           pictures.

Be strict: choose object_visible ONLY if you could actually point at the target
object. Do NOT infer from world knowledge that "eggs are usually in the fridge".

Return ONLY JSON:
{"category": "<one of the four>", "confidence": "high"|"medium"|"low",
  "evidence": "one short sentence naming exactly what you can see"}
```

Instruction-misalignment sweep (instruction-conditioned prediction). Every instruction-conditioned sample was re-judged in a two-stage protocol: a Gemini 3.8 Flash full screen of all 394 samples, then the qwen3.8-flash judge below re-reviewing the newly flagged ones in three repetitions with majority vote, with mocap hard signals arbitrating; stage 1 feeds it the neutral per-view descriptions (Table 18). 51 samples were removed and 14 hand-side mislabels flagged rather than deleted (Sec. C.6).

```
You are auditing a temporal-alignment annotation for an egocentric robot-learning
dataset.

The instruction given to the person (what they are about to do) is:
  "{instr}"

The annotated window is [{t0}, {t1}] seconds. You are shown this exact window from
two cameras, in order:
  1) egocentric (first-person, head-mounted) view | the first {n} frames
  2) exocentric (third-person, room camera) view | the next {n} frames, same
timestamps
Each frame is labeled with its timestamp in seconds.
Also given: neutral event lists describing the two views, written BEFORE seeing the
instruction (unbiased):
  ego: {d_ego}
  exo: {d_exo}

Step 1 | split the instruction into clauses (verb phrases joined by
"and"/"then"/commas, or a single clause; goal modifiers like "until ..." belong to
their clause, they are not separate clauses). For each clause decide, from the frames
AND the event lists, whether that clause's action BEGINS within the window [t0, t1]
(the action itself starting, not merely walking toward or looking at something):
"yes" / "no" / "unclear".
Step 2 | overall status:
  aligned          : every clause begins within the window (a clause may still be in
                     progress at the window end | windows are capped at 6 s)
  misaligned_before: the first clause does not begin in the window because the
                     instructed activity (or its earlier clauses) already happened
                     before the window
  misaligned_after : the instructed action happens only after the window (the window
                     shows unrelated or preparatory activity only)
  different_action : the window shows a clearly different activity than instructed
  unclear          : not enough evidence

Rules:
- Judge the ACTION SEQUENCE, not exact object identity (a bowl seen from above may
look like a cup; a wooden board may look like a tray). Do not fail a window over
object naming alone when the action matches.
- If the egocentric and exocentric evidence disagree, prefer the EGOCENTRIC view for
hand-object interaction detail (it is much closer); use the exocentric view for gross
body position and movement.
- A camera that cannot see the action (person occluded, hands below the counter,
person too small/far) is uninformative | NOT evidence that the action did not happen.
- The neutral event lists can contain errors; verify against the frames.

Reply JSON only:
```

```
{"clauses": ["<clause>", ...],
  "clause_begins": ["yes"|"no"|"unclear", ...],
  "what_happens": "<1-2 lines: what the person actually does in this window>",
  "status":
"aligned"|"misaligned_before"|"misaligned_after"|"different_action"|"misaligned"|"unclear",
  "reason": "<one line>",
  "confidence": "high"|"low"}
```

H.1.6 GROUND-TRUTH REFINEMENT

Both refinement prompts—the per-track refinement of mocap-projected boxes on zoomed crops (qwen3.8-flash) and the GPT-6-astra re-localization of the exocentric reference box from the clean full-resolution frame—are registered in Table 18 (Sec. C.6).

H.2 EVALUATION PROMPTS

All eight tasks are evaluated zero-shot with a single user message and no system message; the text below is the complete instruction the models see, taken from the released manifests. For the four tasks whose instruction is a fixed template (object correspondence—identical in both source batches—HOI correspondence, trajectory grounding, and goal-conditioned prediction), the template is byte-identical across all samples of the task. For the remaining tasks, whose instructions embed per-sample slots (task correspondence's task description, instruction-conditioned prediction's instruction, and the planning questions' goal and options), we quote one real sample verbatim and name its `sample_id`; the slots vary per sample, everything else is fixed.

Object correspondence. Fixed template, identical in the ACE and Ego-Exo4D batches (n=125+200). Sample `take-000002__Water_Cup_White__front_left__zed_cam4__t23.4`.

```
You are given an exocentric (third-person) reference frame in which one object is
marked with a bounding box, and an egocentric (first-person) clip that starts before
and ends after the reference moment. Track the marked object through the egocentric
clip.
Output Format: [{"time": 1.0, "bbox_2d": [x_min, y_min, x_max, y_max], "label": ""},
...] -- times in seconds relative to the start of the egocentric clip; omit
timestamps where the object is not visible.
```

HOI correspondence. Fixed template (n=279). Sample `take-000002__Water_Cup_White__front_left__zed_cam2__t23.4`.

```
You are given two temporally synchronized video clips of the same activity: a short
clip from an exocentric (third-person) reference viewpoint, and a longer clip from
an egocentric (first-person) target viewpoint that starts at or before, and ends at
or after, the reference clip. The exocentric clip shows a single hand-object
interaction. Watch it, then localize the time range within the egocentric clip (in
seconds, relative to the start of the egocentric clip) during which the same
interaction occurs, and track every hand and object involved in hand-object
interactions throughout that time range.
Output Format: {"time_range": [start, end], "hoi_track": [{"time": 1.0,
"object_bbox_2d": [x_min, y_min, x_max, y_max], "object_label": "", "hand_bbox_2d":
[x_min, y_min, x_max, y_max], "hand_label": ""}, ...]}
```

Task correspondence. The task description after "The task:" is the per-sample slot (n=79). Sample `take-000002__task__front_left__zed_cam2`.

```
You are given a reference-view video clip and a single target-view image, both
covering the same task within the same time span. The reference view is a
third-person (exocentric) recording of someone performing a multi-step task; the
target view is a single first-person (egocentric) frame. The task: Task: Pick up a
metal cup from the coffee table, carry the cup to the kitchen sink, rinse the cup,
and place the cup on the counter.
Identify the objects that are interacted with in the reference clip to complete the
task, in the order they are used, and localize each one in the target-view frame.
Output Format: [{"step": 1, "bbox_2d": [x_min, y_min, x_max, y_max], "label": ""},
{"step": 2, "bbox_2d": [x_min, y_min, x_max, y_max], "label": ""}, ...]
If an object is not visible in the target-view frame, set "bbox_2d" to null but
still name it in "label".
```

Next-step and long-horizon planning. The goal text and the four options are per-sample slots over one fixed skeleton shared by both planning tasks, which differ only in the action count—the next 1 versus the next 3 action(s) (n=493+699 pooled). Sample `sfu_cooking028_2_k22_cam_v3`.

Next-step planning:

```
Given the historical task progress and the current observation, what are the next 1
action(s) the camera wearer should take in order to Take out and peel a garlic
clove?

You are given an egocentric (first-person) video clip of the task so far and one
exocentric (third-person) frame. Choose the single best option:
A. pick knife from countertop (on the kitchen island countertop)
B. peel garlic clove (on the blue cutting board near the sink)
C. pick knife from countertop (on the counter to the right of the sink)
D. hold garlic clove (on the granite kitchen island)
```

Long-horizon planning (same skeleton; only the action count and the options differ):

```
Given the historical task progress and the current observation, what are the next 3
action(s) the camera wearer should take in order to Take out and peel a garlic
clove?

You are given an egocentric (first-person) video clip of the task so far and one
exocentric (third-person) frame. Choose the single best option:
A. pull garlic clove from mesh bag (on the granite countertop) -> drop mesh bag on
countertop (on the granite countertop) -> hold garlic clove (in the person's hand
over the granite countertop)
B. pick knife from countertop (on the countertop between the sink and the dish
rack) -> place garlic clove on chopping board (on the blue cutting board in front
of the sink) -> turn garlic clove on cutting board (on the blue cutting board in
front of the sink)
C. pick knife from countertop (in the drying rack next to the toaster) -> place
garlic clove on chopping board (on the granite island counter near the eggs) ->
turn garlic clove on cutting board (on the black dish drying rack)
D. place knife on garlic clove on chopping board (on the blue cutting board on the
counter) -> place knife on countertop (on the granite countertop near the sink) ->
pick garlic clove from chopping board (on the blue cutting board)
```

Trajectory grounding. Fixed template (n=77). Sample `take-000000__front_left__zed_cam2__t83.3`.

```
You are given a video clip from an exocentric (third-person) reference viewpoint,
and a sequence of keyframe images from an egocentric (first-person) target
viewpoint, listed in temporal order (frame 1 is earliest). The keyframes are sampled
at moments when the egocentric camera's viewpoint changes sharply. The reference
clip spans the same time range as the keyframe sequence and shows one or both hands
moving while performing an action. For each keyframe, localize the 2D positions of
the hand-motion trajectory points that occur from that keyframe up to (but not
including) the next one, expressed in that keyframe's own image coordinates. Only
include a hand's points for the timestamps when it is visible in that keyframe's
view; omit it at the other timestamps.
Output Format: [{"frame": 1, "trajectory_2d": [{"point_2d": [x, y], "label":
"left_hand"}, {"point_2d": [x, y], "label": "right_hand"}, ...]}, {"frame": 2, ...},
...]
```

Goal-conditioned trajectory prediction. Fixed template (n=202). Sample `take-000000__Coffee_Cup__front_left__zed_cam1__t51.2`.

```
You are given three images: an egocentric (first-person) current view, and two
exocentric (third-person) views of the same scene -- one captured at the current
moment and one captured later, showing the goal state after an action is completed.
Predict the trajectory of the hands from the current moment until the goal state:
for each timestamp, give the 2D position of each hand in the EGOCENTRIC current
view's image coordinates.
Output Format: [{"time": 0.0, "point_2d": [x, y], "label": "left_hand"}, {"time":
0.0, "point_2d": [x, y], "label": "right_hand"}, {"time": 0.5, ...}, ...] -- times
in seconds relative to the current moment; omit a hand at a timestamp if it is not
visible in the egocentric view.
```

Instruction-conditioned trajectory prediction. The quoted instruction is the per-sample slot (n=343). Sample `take-000000__Coffee_Cup__front_left__zed_cam5__t49.2`.

```
You are given an exocentric (third-person) view and an egocentric (first-person)
view captured at the current moment, and an instruction describing what the person
is about to do: "pick up the cup from the white office table" Predict the
trajectory of the hands from the current moment until the instruction is fulfilled:
for each timestamp, give the 2D position of each hand in the EGOCENTRIC current
view's image coordinates. The coordinate system is fixed at the current moment;
omit a hand at a timestamp if it is not visible in this view.
Output Format: [{"time": 0.0, "point_2d": [x, y], "label": "left_hand"}, {"time":
0.0, "point_2d": [x, y], "label": "right_hand"}, ...] -- times in seconds relative
to the current moment.
```

H.2.1 VIEW ASSEMBLY AND INPUT CONDITIONS

Both evaluation harnesses assemble each request as a single user message whose content parts follow the manifest's view order, media first and the instruction text last (`evaluation/mveb_eval.py`, `build_content`; the second harness mirrors this assembly). A *clip* view becomes one video part sampled client-side on a uniform 2 fps grid under the per-task caps of Sec. E.3; a *frame* view becomes a single image part; a *frames* list becomes consecutive image parts. In the gt-aligned input mode of the official object- and HOI-correspondence calibers, the target clip is instead expanded into one image part per frame on the ground-truth 0.5 s grid, each preceded by a plain-text timestamp part of the exact form `<12.5 seconds>`—the same string the Qwen3-VL video processor renders natively, so per-frame and native video inputs produce identical prompts. Gateways that reject video parts receive the same sampled frames as image parts with the same stamps, and per-model image budgets are applied by uniform subsampling with the first and last frames kept (Sec. E.3).

The viewpoint conditions of Section 4.3 are pure view filters over this assembly: *ego+exo* sends all views and *ego-only* keeps the target-role views; the instruction is byte-identical across conditions and models. Output parsing and scoring for the answer formats above are specified in Sec. E.4; answer schemas and per-task construction details are in Appendix C.