

EgoTools: Towards Tool-Centric Reasoning in Real-World Egocentric Videos

Shulin Tian^{1,2,*}, Junsu Kim^{1,3,*}, Shuai Liu^{1,*}, Hao Li^{1,4,*},
Yujiao Shen¹, Sihan Li¹, Zhe Yang¹, Yeongon Kim¹,
Runmao Yao¹, Yuhao Dong¹, Fangzhou Hong^{1,4}, Antonino Furnari⁵,
Jingkang Yang¹, Hongyuan Zhu², Ziwei Liu^{1,†}

¹S-Lab, Nanyang Technological University

²A*STAR, ³KAIST, ⁴Ropedia, ⁵University of Catania

<https://ropedia.github.io/egotools>

Abstract

Real-world embodied tasks require agents to act under physical constraints while tracking evolving object and task states. Because tools mediate many everyday and professional activities, understanding these tasks requires reasoning about affordances, hand–tool–object geometry, procedural progress, and causal effects. Tool-centric embodied reasoning therefore calls for real-world egocentric training data and diagnostic benchmarks that connect these abilities. We introduce **EgoTools**, the first comprehensive suite for egocentric tool-use understanding. It combines **EgoTools-Data**, 100.36 hours of tool-centric egocentric recordings with synchronized audio, dense captions, reasoning-heavy narrations, and supplementary 3D information, with **EgoTools-Bench**, a diagnostic benchmark of 1,000 QA pairs across four tracks spanning perception, geometry, procedure, and causal reasoning. EgoTools-Bench reveals a substantial gap between current models and human experts. Under strict source-video separation, supervised fine-tuning of Qwen3-VL-8B-Instruct on instruction data derived from EgoTools-Data improves accuracy from 50.0% to 60.9% on the full 1,000-question benchmark. These results establish EgoTools as a unified training and diagnostic evaluation resource for real-world egocentric tool-use understanding.

1 Introduction

Tool use is central to physical activities ranging from daily chores to professional procedures. Understanding these activities requires an embodied agent to recognize actions and objects while reasoning about tool–target contact, spatial coordination, procedural progress, and physical outcomes. Egocentric video foregrounds hands, tools, and their immediate consequences within an evolving task

state, making it a demanding testbed that tightly couples perception with temporal and causal reasoning.

Current multimodal video models (Bai et al., 2025a,b; Google DeepMind, 2026b; Li et al., 2025a; Zhu et al., 2025) perform strongly on perception-oriented video tasks. Existing egocentric datasets (Damen et al., 2022; Grauman et al., 2022, 2024; Liu et al., 2022; Ragusa et al., 2023; Sener et al., 2022) typically organize annotations around activities, objects, or hand–object interactions, and benchmarks (Chen et al., 2026; Cheng et al., 2024; Di and Xie, 2024; Jia et al., 2022; Mangalam et al., 2023; Plizzari et al., 2025) often formulate questions at the activity level, such as “cook eggs.” Tool-centric evaluation adds explicit links among the mediating tools, target-object state, manipulation context, and causal outcomes.

We introduce **EgoTools**, the first comprehensive suite combining training and evaluation resources for egocentric tool-use understanding. **EgoTools-Data** contains 100.36 hours of real-world tool-centric recordings across everyday and professional tasks, with synchronized audio, dense captions, narrations covering affordances, procedure, and causal effects, and supplementary 3D information. These annotations provide structured supervision for tool choice and substitution, object-state changes, procedural progress, and grounded hand–tool–object interactions.

From this corpus, we construct **EgoTools-Bench**, a benchmark of 1,000 question–answer pairs across four tracks. **Affordance & Causality** tests goal-directed tool choice, affordances, preconditions, and consequences; **Perception & Grounding** tests visible identities, attributes, states, and quantities; **Procedural Dynamics** tests tool–action order, step transitions, workspace staging, and manipulation; and **Spatial Reasoning** tests egocentric hand–tool–object geometry, including position, alignment, containment, support, contact, and depth. Together,

*Equal contributions. †Corresponding author. Email: {shulin002, ziwei.liu}@ntu.edu.sg.



Figure 1: **Overview of EgoTools.** EgoTools is an egocentric tool-use suite that spans multiple environments, with dense hierarchical captions and long-sequence 4D object tracking. We also propose **EgoTools-Bench**, a 1,000-question QA benchmark spanning diverse real-world environments and reasoning tasks.

they span perception and geometry, procedure, and causal reasoning.

On the quality-controlled full 1,000-question benchmark, Qwen3-VL-8B-Instruct reaches 50.0%. Under strict source-video separation, fine-tuning it on instruction data from the EgoTools-Data training pool raises accuracy to 60.9%, with gains on Affordance & Causality, Perception & Grounding, and Procedural Dynamics and a decrease on Spatial Reasoning. EgoTools therefore quantifies tool-centric reasoning capabilities and provides actionable supervision for improving them.

2 Related Work

Egocentric video datasets and benchmarks. Egocentric video captures hands, manipulated objects, context, and temporally ordered actions from the actor’s view (Bambach et al., 2015; Fathi et al., 2012; Li et al., 2018; Pirsiavash and Ramanan, 2012). Large-scale datasets support first-person action, object, hand-object, procedural, industrial, and skilled-activity understanding (Damen et al., 2022; Grauman et al., 2022, 2024; Liu et al., 2022; Ragusa et al., 2023; Sener et al., 2022), while recent benchmarks assess long-context QA, grounding, planning, assistance, and first-person reasoning (Chen et al., 2026; Cheng et al., 2024; Di and Xie, 2024; Jia et al., 2022; Mangalam et al., 2023;

Plizzari et al., 2025; Yang et al., 2025; Zhou et al., 2025). These resources largely organize annotations around activities, objects, events, or plans, and instructional-video tasks study detours and step differences involving ingredients, tools, or techniques (Ashutosh et al., 2024; Nagarajan and Torresani, 2024). EgoTools centers tool choice, substitution, and state-dependent manipulation through multi-environment egocentric video, tool-centric narration, focal-tool grounding, and diagnostic QA.

Physical tool-use reasoning. Physical tool use connects function and affordances to goals, target materials, motion constraints, and evolving object states. Robotics studies related questions through task-oriented grasping, tool-flow prediction, affordance-centric manipulation, and egocentric affordance learning (Fang et al., 2018; Li et al., 2025b; Qin et al., 2023; Seita et al., 2023; Yamanobe et al., 2017), often in structured tasks or short interactions. MLLM and VLM evaluations address affordance grounding and physical tool understanding (Huang et al., 2024; Qian et al., 2024; Yu et al., 2025; Zhang et al., 2025b), frequently using static images, 3D scenes, synthetic data, or isolated interactions. EgoTools-Bench complements both lines with continuous first-person evidence of real tools being selected, substituted, and adapted, testing links among visual evidence, functional choices,

Table 1: **Comparison with representative egocentric datasets and benchmarks.** A checkmark indicates that the feature is a central design component;  indicates partial support. Signals describe sensor/video inputs:  = RGB/video,  = 3D/depth/pose,  = inertial signals, and  = gaze; QA text and construction-time narrations are treated as annotations. Non-RGB signals may cover subsets of large datasets. Hours are total video hours or estimates from source-reported clip counts and durations; “—” marks unavailable entries. Embodied interaction benchmarks provide scope context alongside egocentric video corpora.

Dataset / Benchmark	Scene	Video Hours	#QA	Signals	Multi-Env.	Tool-Centric Narration	Narr-Linked Tool Grounding	QA Benchmark	Tool Choice / Substitution	Physical Tool-Use Reasoning
— Egocentric Observation Datasets —										
EPIC-KITCHENS (Damen et al., 2022)	Kitchen	100	—	 						
Ego4D (Grauman et al., 2022)	Real-world	3,670	14.5 K	   						
Ego-Exo4D (Grauman et al., 2024)	Skilled	1,286	—	   						
Assembly101 (Sener et al., 2022)	Assembly	513	—	  						
MECCANO (Ragusa et al., 2023)	Industrial	6.9	—	  						
HOI4D (Liu et al., 2022)	Indoor HOI	44.4	—	 						
— Egocentric Video Reasoning Benchmarks —										
EgoSchema (Mangalam et al., 2023)	Real-world	250+	5,031							
EgoTaskQA (Jia et al., 2022)	Task videos	14	40 K							
EgoPlan-Bench (Chen et al., 2026)	Daily tasks	—	4,939							
EgoIntent (Pan et al., 2026)	Daily tasks	2.9	3,014							
GroundVQA (Di and Xie, 2024)	Long videos	736	303 K							
EgoTempo (Plizzari et al., 2025)	Temporal	6.3	500							
— Embodied Interaction Benchmarks —										
OpenEQA (Majumdar et al., 2024)	Indoor EQA	—	1.6 K	 						
RoboVQA (Sermanet et al., 2024)	Robotics	238	829 K							
RoboCasa365 (Nasiriany et al., 2026)	Sim kitchens	2,200+	—	 						
EgoTools (Ours)	Real-world	100.36	1,000	  						

manipulation steps, and object-state changes.

3 EgoTools Data Suite

We present EgoTools, a real-world egocentric video suite for physically grounded tool-use understanding. EgoTools-Data contains 100.36 hours of synchronized video, audio, and geometry-related signals, augmented with hierarchical dense captions, 2D-grounded tool-centric narrations, and 3D annotations derived from reconstruction and tracking. From source-video-disjoint pools, the suite provides (i) a video instruction-tuning corpus from the training pool and (ii) EgoTools-Bench, a 1,000-question 8-way diagnostic benchmark from a curated 40.34-hour reserved pool. Its 1,000 questions span Affordance & Causality, Perception & Grounding, Procedural Dynamics, and Spatial Reasoning.

3.1 Data Collection and Preprocessing

We record with **HOMIE**¹, a lightweight head-mounted device from Ropedia, Inc. with four synchronized fisheye views, audio, and auxiliary motion/synchronization signals supporting reconstruction, pose/depth estimation, and object/hand tracking. For annotation and training, we rectify the front-left RGB stream into a natural first-person view, with zoom and pitch adjustments

that preserve hand–tool–object interactions (Appendix B.1). The final videos are 1024×1024 , 20 FPS, single-view egocentric streams synchronized with audio.

Collection spans **daily activities** and **expertise-intensive procedures** across seven domains: kitchen, classroom, research lab, repair workshop, craft, office, and household (Figure 1). These cover everyday manipulation, material transformation, manual operations, and educational and professional experiments. Recordings were collected in kitchens, workshops, laboratories, and daily-living spaces at universities and research sites in Asia. Graduate and undergraduate collectors had diverse disciplinary backgrounds; expertise-intensive procedures were performed or reviewed by collectors familiar with the task to ensure natural and technically valid tool use. Appendix A details the domains and anonymized participants.

The curated corpus contains **100.36 hours** across these domains. **Structured task-guided** recordings follow predefined goals and procedural steps, yielding clear boundaries, observable state changes, and controlled hand–tool–object coverage. In **open-ended participant-driven** recordings, participants receive a broad topic or goal and choose how to complete it, yielding spontaneous tool choice, adaptation, repeated attempts, error recovery, and substitution. The two settings combine consistent procedural coverage with in-the-wild variability.

¹https://ropedia.com/blog/20251216_introducing_ropedia

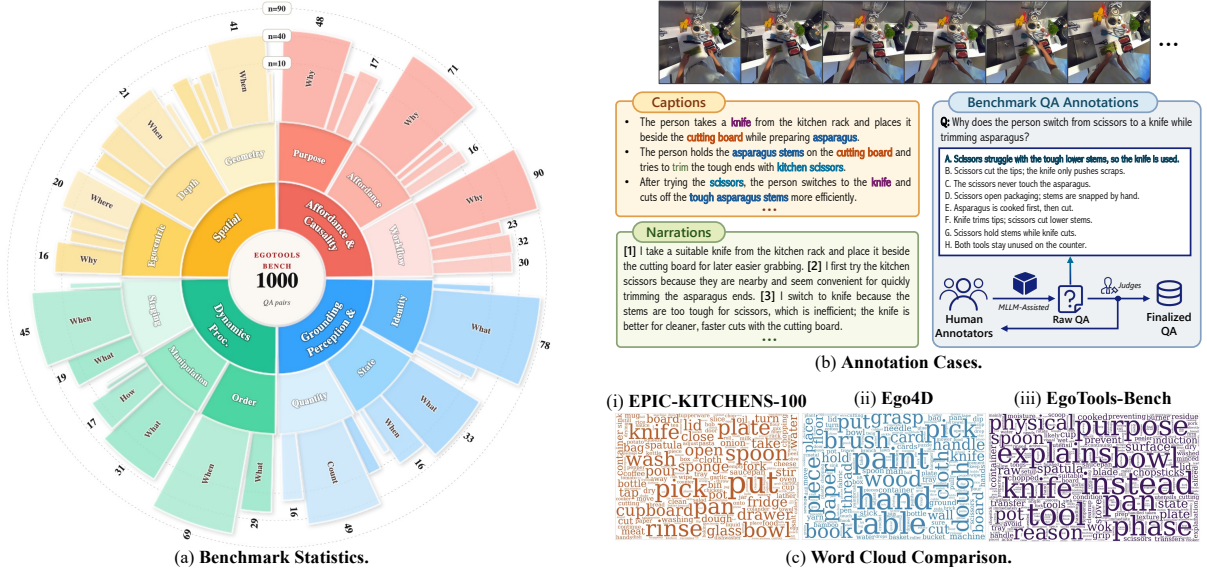


Figure 2: **Benchmark distribution and annotation examples.** (a) Distribution of 1,000 QA pairs across four tool-use reasoning tracks and their fine-grained subtracks. (b) Example annotations, including captions, tool-centric narrations, and finalized QA pairs produced through human annotation with MLLM-assisted checking. (c) Word-cloud comparison with EPIC-KITCHENS-100 and Ego4D, showing that EgoTools-Bench is more tool-dense and centered on tools, actions, objects, and state changes.

Source-video partition. We partition source videos into training and benchmark-reserved pools before constructing either resource. Instruction data is derived exclusively from the training pool, while the 40.34-hour reserved pool supplies the benchmark clips and annotations.

3.2 Multimodal Annotation and Curation

Textual Annotations. We provide **dense captions** for visible actions and task progress and **tool-centric narrations** for intent, grounding, and reasoning. A streaming hierarchical Gemini-3-Flash (Google DeepMind, 2025) pipeline splits each video into 5-minute clips and uses 32 uniformly sampled frames per clip for a global context caption. It then captions sequential 5-second chunks, conditioning the first on global context and later chunks on the preceding caption, and aggregates them into 1-minute windows and 5-minute summaries.

Key-frame object recognition supplies evidence for post-hoc hallucination checks. Following the use of narration as weak supervision in egocentric datasets (Damen et al., 2022; Grauman et al., 2022), task performers narrate meaningful tool-use events with a template emphasizing tool selection, intent, and effects on target objects or task states. For selected narrations, they mark 2D grounding points on keyframes, linking focal tool or object mentions to visual instances and initializing long-horizon 3D

tracking. Dense captions and corrected English narrations become temporally grounded instruction examples for action description, procedural summarization, tool selection and substitution, object-state tracking, and next-step prediction. Figure 3 shows the interfaces; Appendix B.3 provides templates and examples.

3D Annotations. HOMIE provides egocentric video, camera poses, and depth maps. Following Holi-Spatial (Gao et al., 2026b), we train a 3D Gaussian Splatting model (Li et al., 2026b) for multi-view-consistent geometry and temporally coherent per-frame depth while reducing artifacts such as floaters (Figure 4).

For long-term tracking, tool-centric narrations provide initial 2D points and semantic captions, and SAM2 (Ravi et al., 2025) propagates masks through the video. A VLM (Google DeepMind, 2026b) checks consistency at later grounding points; detected drift or semantic mismatch triggers SAM2 re-initialization and refinement from the failure point. This recursive process yields geometrically precise, semantically coherent spatio-temporal masks.

We lift tracked masks into 3D with reconstructed depth and camera poses to obtain object boxes and long-horizon 4D trajectories, then use spatial QA templates to query motion and relations. The original 100.36-hour videos, textual annotations, and 3D annotations form EgoTools-Data, support-

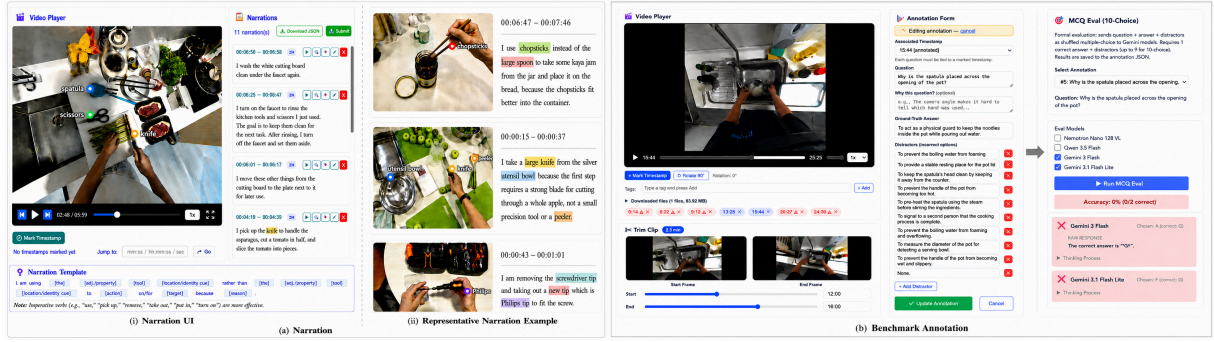


Figure 3: **Textual annotation pipeline.** (a) Tool-centric narrations are annotated from grounded egocentric video segments. (b) Benchmark QA pairs are constructed and refined through an annotation interface with MLLM-assisted checking.

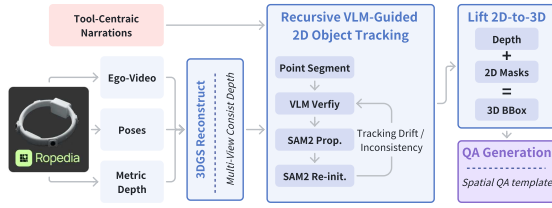


Figure 4: **3D annotation pipeline.** Raw geometry from the HOMIE device is reconstructed with 3DGS to obtain consistent depth, then combined with tool-centric narrations for recursive VLM-guided object tracking. The resulting masks are lifted to 3D for object boxes and spatial QA generation.

ing video instruction tuning and the source-video-disjoint EgoTools-Bench evaluation set.

3.3 Benchmark Construction and Quality Control

Benchmark QA Construction. We select a **40.34-hour** benchmark-reserved subset prioritizing annotation quality, tool-use density, task diversity, and environmental coverage; it is source-video-disjoint from instruction tuning. Original collectors contribute task-aware procedural and causal questions, while external annotators identify visually inferable, model-challenging reasoning points. Domain-specific videos, including wet-lab procedures, are assigned to annotators with relevant expertise.

Annotators identify segments involving tool selection, object-state changes, or temporal ordering and write 8-way questions with one correct answer and seven distractors. Clips are localized around an anchor moment and mostly last at least three minutes. Limiting temporal overlap and per-video QA density maintains diversity across videos, tools, and tasks. Together with the spatially grounded questions described below, this process yields **1,000** benchmark pairs.

The benchmark includes human-verified spatial QA pairs generated from 3D annotations using the Holi-Spatial (Gao et al., 2026b) pipeline, augmented with focal-tool trajectories and object-state signals to cover motion, state changes, and interaction dynamics. The average temporal span is **4.19 minutes** per QA.

Quality Control. All items undergo fix-first quality control that repairs valid annotator intent before removal. We check structural validity, video grounding, and answer choices for malformed questions, empty or duplicate submissions, missing evidence, temporal misalignment, inconsistent tool identity, and weak distractors. Empty or verbatim-duplicate items are removed; repairable reference errors, first-person leakage, and incomplete answers are fixed with deterministic rules or minimal model-assisted edits; low-confidence cases go to human review.

After 8-way normalization, deterministic, text-only, and multi-model checks flag near-duplicate choices, revealing wording, lexical imbalance, answer-length or position bias, and shortcut-solvable items. Reviewers inspect flagged items and replace distractors with visually plausible, length-balanced alternatives while preserving the question and ground truth when possible. Final decisions use visual grounding, unique answerability, and annotation quality as model-independent criteria (Appendix D).

Independent Reliability Study. We independently evaluated a stratified sample of 250 questions, including 100 Affordance & Causality items. Five annotators independent of question authoring each assessed 100 assigned items, giving two independent judgments per question and 500 total. Using the video, question, and choices with

the answer key hidden, annotators agreed with the key in **89.8%** of judgments, including 86.5% for Affordance & Causality. Pairwise exact answer agreement was **88.4%**; **94.4%** of items had a uniquely best visually supported answer. Fourteen items (**5.6%**) were flagged as ambiguous or multiply plausible and manually reviewed; the review revised or replaced affected items to meet the unique-answer criterion. We recomputed all reported results on the resulting quality-controlled benchmark.

3.4 Distribution

Figure 2 summarizes **1,000 QA pairs**: **1) Affordance & Causality (363)**, **2) Perception & Grounding (236)**, **3) Procedural Dynamics (222)**, and **4) Spatial Reasoning (179)**. The tracks span tool recognition, object-state grounding, relations among hands, tools, and objects, task progress, tool selection, application, and effects. Fine-grained subtracks support detailed analysis.

The seven domains characterize the full EgoTools-Data corpus. EgoTools-Bench follows a long-tailed domain distribution, with most questions from Kitchen videos; supplementary material provides complete statistics. We therefore focus the analysis on reasoning capabilities across the full benchmark.

We limit clip overlap and per-video QA density to maintain diversity across recordings. Figure 2 compares vocabulary with EPIC-KITCHENS-100 and Ego4D, qualitatively showing greater emphasis on tools, manipulated objects, actions, and state changes than on general egocentric activity recognition.

4 Experiments

4.1 Experimental Setup

We evaluate human experts, proprietary models, and open-source models on EgoTools-Bench, alongside available comparable results for EgoSchema (Mangalam et al., 2023), EgoPlan-Bench (Chen et al., 2026), and EgoThink (Cheng et al., 2024). Models include standard video-language, audio-capable omni, and thinking variants from Gemini (Google DeepMind, 2026b), Qwen (Bai et al., 2025a,b; Xu et al., 2025; Wang et al., 2024), InternVL (Wang et al., 2025; Zhu et al., 2025), LLaVA (Li et al., 2025a; Zhang et al., 2025a), MiMo-VL (Xiaomi LLM-Core Team et al., 2025), and GLM (GLM-V Team et al., 2026). We

report overall and four-track accuracy on EgoTools-Bench, following the tracks introduced in Section 3.4.

We report frame count and audio use. Most open-source instruct models use 64 uniformly sampled frames, thinking models use 64 or 512 as supported, and Gemini (Google DeepMind, 2026b) uses 1 FPS video. Vision-based models receive sampled frames, and audio-enabled models additionally receive synchronized ambient task audio. Benchmark construction uses corrected narrations, dense captions, and annotation metadata.

Human evaluation. Five evaluators independent of question authoring each answered all 1,000 items over multiple sessions, freely seeking and replaying video with the answer key hidden. All 5,000 judgments were complete and valid. Overall and track-wise accuracies are micro-averaged over these judgments under a free-viewing protocol; model results use their reported fixed-frame budgets.

EgoTools training. We adapt Qwen3-VL-8B-Instruct in a single stage of supervised fine-tuning that updates every parameter in its language-model component; the visual encoder and multimodal aligner retain their pretrained weights throughout. All 184,679 training examples come from the EgoTools-Data training pool, which is source-video-disjoint from EgoTools-Bench. The mixture covers tool-centric reasoning, episode-level multiple-choice and general video QA, dense captioning, narration completion, and single-image open-ended QA. Figure 5 visualizes the mixture, and Table 7 gives the full breakdown. External evaluation uses EgoSchema, EgoPlan-Bench, and EgoThink. Supplementary material provides the optimization details.

4.2 Main Results

EgoTools-Bench measures a substantial model-human performance difference. In Table 2, 8-way chance accuracy is 12.5%, open-source instruct models range from 38.9% to 50.0%, and human experts reach 83.2%. Qwen3-VL-8B-Instruct (Bai et al., 2025a) leads instruct models at 50.0%, followed by InternVL3.5-8B-Instruct (Wang et al., 2025) at 48.7%; GLM4.1V-Thinking (GLM-V Team et al., 2026) leads open-source thinking models at 50.7%. This 32.5-point difference between the strongest open-source zero-shot result and human experts quantifies headroom

Table 2: **Performance comparison across egocentric video understanding benchmarks.** We compare human performance, proprietary models, and open-source video-language models on three existing egocentric benchmarks and EgoTools-Bench. The last five columns report results on EgoTools-Bench across four research-facing tracks and the overall average. Track abbreviations are: **AC** = Affordance & Causality, **PG** = Perception & Grounding, **PD** = Procedural Dynamics, and **SR** = Spatial Reasoning. EgoTools-8B is trained on EgoTools-derived instruction data and is therefore reported separately from zero-shot baselines. Rank highlights cover the open-source and EgoTools rows; human and Gemini rows provide reference points. Dark and light green denote the best and second-best result in each column; underlining denotes third-best, with ties marked equally.

Model	Frames	Audio	EgoSchema	EgoPlan-Bench	EgoThink	EgoTools				
						AC	PG	PD	SR	Overall
Human Baseline										
Human Expert	–	✓	~76	–	–	82.1	85.2	83.3	82.7	83.2
Proprietary Models										
Gemini-3.1-Pro (Google DeepMind, 2026b)	1fps	✓	–	–	–	69.7	51.7	72.1	74.9	66.9
Gemini-3-Flash (Google DeepMind, 2025)	1fps	✓	–	–	–	64.5	50.0	73.0	72.6	64.4
Gemini-3.1-Flash-Lite (Google DeepMind, 2026a)	1fps	✓	–	–	–	58.1	44.5	59.9	58.7	55.4
Open-Source Models (Instruct)										
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	64	✗	69.0	42.3	61.0	46.1	45.9	57.1	55.4	50.0
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	64	✗	69.6	42.0	61.9	41.7	45.4	50.2	50.0	45.7
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	64	✗	61.0	32.0	57.7	40.1	43.3	50.2	47.8	44.3
Qwen2.5-Omni-7B-Instruct (Xu et al., 2025)	64	✓	65.2	–	58.9	44.7	45.9	51.8	46.7	47.1
Qwen2-VL-7B-Instruct (Wang et al., 2024)	64	✗	63.6	39.9	60.1	46.6	35.6	46.2	47.8	44.2
InternVL3.5-8B-Instruct (Wang et al., 2025)	64	✗	63.8	–	56.0	45.8	45.9	53.4	53.3	48.7
InternVL3-8B-Instruct (Zhu et al., 2025)	64	✗	69.6	–	59.4	44.7	44.9	51.8	48.9	47.1
LLaVA-OneVision-7B (Li et al., 2025a)	64	✗	63.4	–	54.9	37.6	37.1	42.9	37.0	38.9
LLaVA-Video-7B-Qwen2 (Zhang et al., 2025a)	64	✗	52.4	–	58.0	38.2	39.7	43.3	37.0	39.8
MiMo-VL-7B-SFT (Xiaomi LLM-Core Team et al., 2025)	64	✗	54.4	35.9	60.7	41.4	44.9	49.4	40.2	44.2
Open-Source Models (Thinking)										
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	512	✗	70.3	43.6	62.0	43.6	41.8	48.2	55.4	45.7
Qwen3-VL-4B-Thinking (Bai et al., 2025a)	512	✗	59.8	43.1	62.3	37.6	36.6	36.4	41.3	37.4
MiMo-VL-7B-RL (Xiaomi LLM-Core Team et al., 2025)	64	✗	57.0	36.6	59.4	43.1	45.9	50.6	39.1	45.3
GLM4.1V-Thinking (GLM-V Team et al., 2026)	64	✗	66.1	25.3	62.7	46.1	50.0	57.1	53.3	50.7
Ours										
EgoTools-8B (Ours)	64	✗	67.6	35.2	64.0	55.4	59.7	68.1	44.4	60.9

in tool-mediated reasoning.

EgoTools-Bench adds tool-specific diagnostic coverage. Qwen3-VL-8B-Instruct reaches 69.0% on EgoSchema (Mangalam et al., 2023), 61.0% on EgoThink (Cheng et al., 2024), and 50.0% on EgoTools-Bench; its 4B counterpart scores 69.6%, 61.9%, and 45.7%, respectively. These cross-benchmark profiles reflect EgoTools-Bench’s additional emphasis on tool affordances, grounding, procedure, spatial relations, and state changes.

Model and input capacity yield heterogeneous gains. Model parameter scale, audio support, and thinking-style inference produce mixed changes in overall accuracy. Their track-specific effects motivate both track-wise analysis and controlled input ablations.

4.3 Training Utility of EgoTools-Data

EgoTools supervision improves tool-centric reasoning. Table 2 shows that supervised fine-tuning on 184,679 EgoTools-derived examples improves Qwen3-VL-8B-Instruct from **50.0% to 60.9%** on EgoTools-Bench. It gains 9.3 points on Affordance & Causality, 13.8 on Perception & Grounding, and 11.0 on Procedural Dynamics, while Spa-

tial Reasoning decreases by 11.0 points. Source-video separation assigns instruction data and evaluation videos to distinct source pools. These results demonstrate the utility of EgoTools-Data as tool-centric supervision.

Existing-benchmark results characterize answer-format coverage. The mixture includes 5,000 next-action multiple-choice and 5,000 single-image open-ended examples, matching EgoPlan-Bench and EgoThink answer formats. EgoTools-8B reaches 64.0% on EgoThink compared with 61.0% for the backbone and 35.2% on EgoPlan-Bench compared with 42.3% for the backbone. These paired results characterize answer-format coverage; our primary claim concerns EgoTools-Data’s training utility for tool-centric reasoning.

4.4 Analysis

Precise visual grounding defines the largest opportunity for improvement. Perception & Grounding requires evidence about the focal tool, target object, contact point, attributes, quantities, and local state changes. Open-source instruct models score 35.6–45.9%. Gemini-3.1-Pro reaches

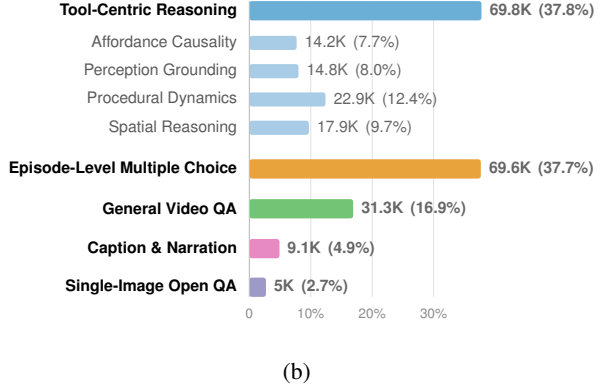
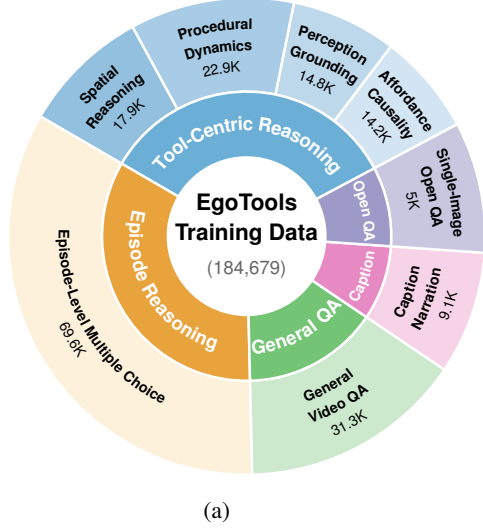


Figure 5: **Composition of the EgoTools-Data instruction-tuning corpus.** (a) The inner ring groups the 184,679 examples by what they supervise, and the outer ring gives the corresponding categories; readability-adjusted segment angles preserve the smallest category labels. (b) Shares drawn to scale; the four tool-centric categories are nested under their group total. Exact counts are listed in Table 7.

Table 3: **Input-evidence ablation on the full 1,000-question EgoTools-Bench.** All conditions use the same Qwen3-VL-8B-Instruct backbone and pre-extracted frame-list evaluation pipeline. Results support within-table comparisons; the main benchmark uses a separate MP4-based protocol.

Input condition	Overall	AC	PG	PD	SR
Ordered 64 frames	51.7	47.4	50.0	57.1	57.6
Shuffled 64 frames	49.3	45.5	47.4	53.9	56.5
Middle frame	46.9	44.7	45.9	51.0	45.7
Text-only	42.7	39.0	36.1	51.4	47.8

51.7% on this track, its lowest track score. Together with Table 3, these results identify precise visual grounding as a priority for future progress.

Model designs produce distinct performance profiles. Scale, audio, and thinking-style inference have heterogeneous track-wise effects. Qwen3-VL-8B-Instruct shows its largest gains over the 4B counterpart on Procedural Dynamics and Spatial Reasoning; Qwen3-VL-8B-Thinking scores lower than its instruct counterpart on AC, PG, and PD. GLM4.1V-Thinking achieves the strongest open-source zero-shot overall score of 50.7%; Appendix E provides further track-wise and cue-based analyses.

4.5 Input-Evidence Ablation

Table 3 compares evidence conditions within a matched frame-list evaluation pipeline. Ordered 64-frame input reaches 51.7% overall accuracy, compared with 49.3% for shuffled frames, 46.9% for a

	Qwen3-8B Inst.	Qwen3-8B Think	InternVL3.5 8B	Qwen2.5 Omni	GLM4.1V Think
Qwen3-VL-8B Inst. 50.0%	1.00	0.59	0.59	0.55	0.55
Qwen3-VL-8B Think 45.7%	0.59	1.00	0.55	0.52	0.54
InternVL3.5-8B Inst. 48.7%	0.59	0.55	1.00	0.54	0.53
Qwen2.5-Omni-7B 47.1%	0.55	0.52	0.54	1.00	0.51
GLM4.1V Think 50.7%	0.55	0.54	0.53	0.51	1.00

Figure 6: **Jaccard overlap of model errors on EgoTools-Bench.**

single middle frame, and 42.7% for text-only input. Perception & Grounding obtains the largest visual-input gain, improving from 36.1% with text only to 50.0% with ordered frames. Procedural Dynamics shows the largest ordered-versus-shuffled difference at 3.2 points, directly demonstrating the value of temporal order. Detailed protocol interpretation is provided in Appendix E.1.

4.6 Error-Overlap Analysis

Error overlap reveals systematic failure patterns. The error map in Figure 6 compares the Jaccard overlap between models’ incorrect predictions. Qwen3-VL-8B-Instruct has the largest overlap with InternVL3.5-8B-Instruct (0.592), closely followed by Qwen3-VL-8B-Thinking (0.586). The

smallest overlap is between Qwen2.5-Omni-7B-Instruct and GLM4.1V-Thinking (0.513). Because raw Jaccard overlap also increases with the models’ marginal error rates, we compare these values with the overlap expected if each pair failed independently. The independence baselines range from 0.330 to 0.366, and every observed overlap is substantially higher; chance-normalized excess overlap ranges from 0.250 to 0.382. Thus, the evaluated architectures share a substantial core of difficult questions, while differences among their remaining errors leave room for complementary modeling improvements.

Taken together, the track-wise results, implicit-explicit ablation, and error-overlap analysis establish EgoTools-Bench as a capability-specific diagnostic. It measures visual grounding of tool-mediated interactions, reasoning about tool choice and state change, and dependence on language, task-order, or commonsense priors.

5 Conclusion

EgoTools combines real-world tool-use recordings, dense annotations, spatial grounding, and diagnostic QA. EgoTools-Bench quantifies current models’ ability to reason about tool-mediated action, causality, procedure, and state change, while supervised fine-tuning on EgoTools-derived instruction data improves Qwen3-VL-8B-Instruct from 50.0% to 60.9% on EgoTools-Bench. The 83.2% human score and track-wise analysis identify precise visual grounding as a key direction for future progress. By providing both diagnostic evaluation and actionable supervision, EgoTools supports progress toward embodied systems that better observe, reason about, and assist with real-world tool use.

6 Limitations

EgoTools-Data spans seven real-world domains, with current coverage concentrated in the represented tools, environments, participant populations, cultures, and professional workflows. EgoTools-Bench follows the available recordings and has a long-tailed domain distribution concentrated in Kitchen videos. Its clips include occlusion, rapid motion, and small objects, yielding sparse visual evidence in challenging cases. Grounding annotations, narration audio, and 3D trajectory assets enter the public release after annotation finalization and applicable authorization and privacy review.

7 Ethics

EgoTools is constructed from voluntarily contributed egocentric recordings, narrations, and annotations under informed consent covering collection, research use, and release. The public release contains anonymized, privacy-reviewed materials, and project privacy controls govern raw fish-eye recordings and sensitive content. The benchmark supports aggregate model research on tool-use reasoning, with dataset documentation defining participant-privacy and intended-use boundaries. Appendix H provides further safeguards and intended-use details.

8 Acknowledgments

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002). This research is also supported by cash and in-kind funding from NTU S-Lab and industry partner(s).

References

- Kumar Ashutosh, Zihui Xue, Tushar Nagarajan, and Kristen Grauman. 2024. [Detours for navigating instructional videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18804–18815.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-VL technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-VL technical report](#). *Preprint*, arXiv:2502.13923.
- Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. 2015. [Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions](#). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957.
- Leonard Bärman and Alex Waibel. 2022. [Where did i leave my keys? — episodic-memory-based question answering on egocentric videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1560–1568.

- Eun Chang, Zhuangqun Huang, Yiwei Liao, Sagar Ravi Bhavsar, Amogh Param, Tammy Stark, Adel Ahmadyan, Xiao Yang, Jiaqi Wang, Ahsan Abdullah, Giang Nguyen, Akil Iyer, David Patrick Hall, Elissa Li, Nicolas Scheffer, Ahmed Kirmani, Babak Damavandi, Rakesh Wang, Anuj Kumar, and 3 others. 2025. [WearVQA: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 164901–164920.
- Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2026. [EgoPlan-Bench: Benchmarking multimodal large language models for human-level planning](#). *International Journal of Computer Vision*, 134(3):118.
- Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. [Ego-Think: Evaluating first-person perspective thinking capability of vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14291–14302.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. [Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100](#). *International Journal of Computer Vision*, 130(1):33–55.
- Shangzhe Di and Weidi Xie. 2024. [Grounded question-answering in long egocentric videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12943.
- Kuan Fang, Yuke Zhu, Animesh Garg, Andrey Kurenkov, Viraj Mehta, Fei-Fei Li, and Silvio Savarese. 2018. [Learning task-oriented grasping for tool manipulation from simulated self-supervision](#). In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania.
- Alireza Fathi, Yin Li, and James M. Rehg. 2012. [Learning to recognize daily actions using gaze](#). In *Computer Vision – ECCV 2012*, pages 314–327. Springer Berlin Heidelberg.
- Hengjian Gao, Kaiwei Zhang, Shibo Wang, Mingjie Chen, Qihang Cao, Xianfeng Wang, Yucheng Zhu, Xiongkuo Min, Wei Sun, Dandan Zhu, and Guangtao Zhai. 2026a. [LifeEval: A multimodal benchmark for assistive AI in egocentric daily life tasks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 32892–32902.
- Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2024. [Physically grounded vision-language models for robotic manipulation](#). In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469.
- Yuanyuan Gao, Hao Li, Yifei Liu, Xinhao Ji, Yuning Gong, Yuanjun Liao, Fangfu Liu, Manyuan Zhang, Yuchen Yang, Dan Xu, Xue Yang, Huaxi Huang, Hongjie Zhang, Ziwei Liu, Xiao Sun, Dingwen Zhang, and Zhihang Zhong. 2026b. [Holi-Spatial: Evolving video streams into holistic 3D spatial intelligence](#). In *Forty-third International Conference on Machine Learning*. Spotlight.
- GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. [GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning](#). *Preprint*, arXiv:2507.01006.
- Google DeepMind. 2025. Gemini 3 Flash model card. <https://deepmind.google/models/model-cards/gemini-3-flash>. Published December 2025.
- Google DeepMind. 2026a. Gemini 3.1 Flash-Lite - model card. <https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/>. Published 3 March 2026.
- Google DeepMind. 2026b. Gemini 3.1 Pro - model card. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>. Published 19 February 2026.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. [Ego4D: Around the world in 3,000 hours of egocentric video](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012.
- Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, and 81 others. 2024. [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 (CVPR)*, pages 19383–19400.
- Siyuan Huang, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoqi Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. 2024. [ManipVQA: Injecting robotic affordance and physically grounded information](#).

- into multi-modal large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7580–7587.
- Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. 2022. **EgoTaskQA: Understanding human tasks in egocentric videos**. In *Advances in Neural Information Processing Systems*, volume 35, pages 3343–3360. Curran Associates, Inc.
- Kangsan Kim, Yanlai Yang, Suji Kim, Woongyeong Yeo, Youngwan Lee, Mengye Ren, and Sung Ju Hwang. 2026. **MA-EgoQA: Question answering over egocentric videos from multiple embodied agents**. *Preprint*, arXiv:2603.09827.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. **LLaVA-OneVision: Easy visual task transfer**. *Transactions on Machine Learning Research*.
- Chuhan Li, Rilyn Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. 2026a. **SAW-Bench: Learning situated awareness in the real world**. *Preprint*, arXiv:2602.16682.
- Gen Li, Nikolaos Tsagkas, Jifei Song, Ruairidh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. 2025b. **Learning precise affordances from egocentric videos for robotic manipulation**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10581–10591.
- Hao Li, Minghan Qin, Zhengyu Zou, Diqi He, Xinhao Ji, Bohan Li, Bingquan Dai, Dingwen Zhang, and Junwei Han. 2026b. **LangSurf: Language-embedded surface gaussians for 3D scene understanding**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13.
- Yin Li, Miao Liu, and James M. Rehg. 2018. **In the eye of beholder: Joint learning of gaze and actions in first person video**. In *Computer Vision – ECCV 2018*, pages 639–655. Springer International Publishing.
- Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2022. **HOI4D: A 4D egocentric dataset for category-level human-object interaction**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022.
- Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, and 5 others. 2024. **OpenEQA: Embodied question answering in the era of foundation models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. **EgoSchema: A diagnostic benchmark for very long-form video language understanding**. In *Advances in Neural Information Processing Systems*, volume 36, pages 46212–46244. Curran Associates, Inc.
- Tushar Nagarajan and Lorenzo Torresani. 2024. **Step differences in instructional video**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18740–18750.
- Soroush Nasiriany, Sepehr Nasiriany, Abhiram Mad-dukuri, and Yuke Zhu. 2026. **RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots**. In *The Fourteenth International Conference on Learning Representations*.
- Ye Pan, Chi Kit Wong, Yuanhuiyi Lyu, Hanqian Li, Chenfei Liao, Jiahao Huo, Lutao Jiang, Zixin Zhang, Jiacheng Chen, Yuqian Fu, and Xu Zheng. 2026. **EgoIntent: A pre-outcome micro-step benchmark for understanding what, why, and next**. *Preprint*, arXiv:2603.12147.
- Hamed Pirsiavash and Deva Ramanan. 2012. **Detecting activities of daily living in first-person camera views**. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854. IEEE.
- Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. 2025. **Omnia de EgoTempo: Benchmarking temporal understanding of multi-modal LLMs in egocentric videos**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24129–24138.
- Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. 2024. **AffordanceLLM: Grounding affordance from vision language models**. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7587–7597.
- Meiying Qin, Jake Brawer, and Brian Scassellati. 2023. **Robot tool use: A survey**. *Frontiers in Robotics and AI*, 9:1009488.
- Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. 2023. **MECCANO: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain**. *Computer Vision and Image Understanding*, 235:103764.
- Francesco Ragusa, Michele Mazzamuto, Rosario Forte, Irene D’Ambra, James Fort, Jakob Engel, Antonino Furnari, and Giovanni Maria Farinella. 2026. **Ego-EXTRA: video-language egocentric dataset for EXpert-TRAinee assistance**. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4438–4450.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura

- Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2025. [SAM 2: Segment anything in images and videos](#). In *The Thirteenth International Conference on Learning Representations*. Oral presentation.
- Daniel Seita, Yufei Wang, Sarthak J. Shetty, Edward Yao Li, Zackory Erickson, and David Held. 2023. [ToolFlowNet: Robotic manipulation with tools via predicting tool flow from point clouds](#). In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1038–1049. PMLR.
- Fadime Sener, Dibyadip Chatterjee, Daniel Sheleпов, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. [Assembly101: A large-scale multi-view video dataset for understanding procedural activities](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21096–21106.
- Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwivedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J. Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Panag Sanketi, Karol Hausman, Izhak Shafran, and 2 others. 2024. [RoboVQA: Multimodal long-horizon reasoning for robotics](#). In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652.
- Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. 2025. [Visual intention grounding for egocentric assistants](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2512–2522.
- Qiance Tang, Ziqi Wang, Jieyu Lin, Ziyun Li, Barbara De Salvo, and Sai Qian Zhang. 2026. [EgoEverything: A benchmark for human behavior inspired long context egocentric video understanding in AR environment](#). *Preprint*, arXiv:2604.08342.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. [InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *Preprint*, arXiv:2508.18265.
- Junbin Xiao, Shenglang Zhang, Pengxiang Zhu, and Angela Yao. 2026. [Ego-grounding for personalized question-answering in egocentric videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40537–40547.
- Xiaomi LLM-Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. [MiMo-VL technical report](#). *Preprint*, arXiv:2506.03569.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. [Qwen2.5-Omni technical report](#). *Preprint*, arXiv:2503.20215.
- Natsuki Yamanobe, Weiwei Wan, Ixchel G. Ramirez-Alpizar, Damien Petit, Tokuo Tsuji, Shuichi Akizuki, Manabu Hashimoto, Kazuyuki Nagata, and Kensuke Harada. 2017. [A brief review of affordance in robotic manipulation research](#). *Advanced Robotics*, 31(19–20):1086–1101.
- Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhonggang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, and 3 others. 2025. [EgoLife: Towards egocentric life assistant](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28885–28900.
- Lei Yao, Yong Chen, Yuejiao Su, Yi Wang, Moyun Liu, and Lap-Pui Chau. 2026. [HAMMER: Harnessing MLLMs via cross-modal integration for intention-driven 3D affordance grounding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23805–23815.
- Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibe Yang, Jingyi Yu, and Jingya Wang. 2025. [SeqAfford: Sequential 3D affordance reasoning via multimodal large language model](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1691–1701.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025a. [LLaVA-Video: Video instruction tuning with synthetic data](#). *Transactions on Machine Learning Research*.
- Zixin Zhang, Kanghao Chen, Xingwang Lin, Lutao Jiang, Xu Zheng, Yuanhuiyi Lyu, Litao Guo, Yinchuan Li, and Ying-Cong Chen. 2025b. [PhysToolBench: Benchmarking physical tool understanding for MLLMs](#). *Preprint*, arXiv:2510.09507.
- Wenqi Zhou, Kai Cao, Hao Zheng, Yunze Liu, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio W. Mayol-Cuevas, Fan Zhang, Weizhe Lin, and Junxiao Shen. 2025. [X-LeBench: A benchmark for extremely](#)

long egocentric video understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15206–15222, Suzhou, China. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, and 35 others. 2023. [RT-2: Vision-language-action models transfer web knowledge to robotic control](#). In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.

A Dataset Collection and Contributor Information

A.1 Data Collection Settings

EgoTools-Data uses two complementary collection settings to balance procedural coverage with natural variation in real-world tool use. In the **structured task-guided setting**, the collection team specifies a task goal and a sequence of key procedural steps while allowing participants to execute each manipulation naturally rather than following frame-level scripts. For example, a kitchen recording may involve washing ingredients, cutting them on a board, heating a pan, manipulating food with appropriate utensils, and transferring the result to a plate. This setting provides clear task boundaries, observable task-state transitions, and repeated hand-tool-object interactions while preserving natural execution.

In the **open-ended participant-driven setting**, participants receive only a broad topic or high-level task goal and determine their own procedure. These recordings capture spontaneous tool selection, alternative strategies, repeated attempts, error recovery, and opportunistic tool substitution, thereby introducing greater variation in action ordering and tool-use strategy. Together, the two settings provide complementary supervision: structured recordings support consistent procedural annotation, while open-ended recordings preserve the variability and adaptability of naturally occurring tool-mediated behavior.

The corpus spans seven tool-use domains: kitchen, classroom, research laboratory, repair workshop, craft, office, and household. Recordings were collected across kitchens, workshops, laboratories, and daily-living spaces at universities and research sites in Asia. For expertise-intensive procedures, recording or review was assigned to contributors familiar with the corresponding task whenever specialized knowledge was required, helping ensure that the captured tool use was both natural and technically valid.

A.2 Anonymous Contributor Profiles

To preserve contributor privacy, we report contributor information only in aggregate and coarse-grained form. We do not disclose names, contact information, institution-specific identifiers, public profiles, or exact per-video contribution counts. Contributors may participate in multiple roles; therefore, the role counts in Table 4 are not mu-

Table 4: **Summary of anonymous contributor roles.** Contributors may participate in multiple roles.

Role	# Contributors
Participant / narrator	16
QA annotator	10
QA reviewer	2
Data curator	3
Grounding annotator	1
Trajectory annotator	2

Table 5: **Aggregate contributor backgrounds.** Statistics are reported only in aggregate and are not linked to individual identities or released recordings.

Degree		Field	
Undergraduate	12	Computer Science	12
Master’s	5	Electrical Engineering	7
PhD	5	Biology	2
Postdoc	1	Chemistry	1
		Film / Media	1
Age Group			
18–21	4	22–24	10
25–27	6	28+	3

tually exclusive. These roles encompass egocentric recording, tool-centric narration, human-crafted benchmark QA construction, independent QA review, dataset curation, focal-tool grounding, and trajectory annotation.

Contributor backgrounds span computer science, electrical engineering, biology, chemistry, and film/media. Specialized procedures, particularly laboratory and technical tasks, were recorded or annotated by contributors with relevant familiarity whenever possible. These statistics document the range of backgrounds represented in EgoTools-Data without linking released recordings to individual identities. Any internal mapping required for data management is kept separate from the public release.

B Preprocessing and Annotation Protocols

B.1 View Rectification Settings

EgoTools-Data is recorded with HOMIE, a head-mounted multimodal recording device that captures four synchronized fisheye camera streams together with audio and auxiliary motion and synchronization signals. These raw modalities support downstream view rectification, temporal alignment, and selected spatial processing. For annotation, benchmark construction, and model training,

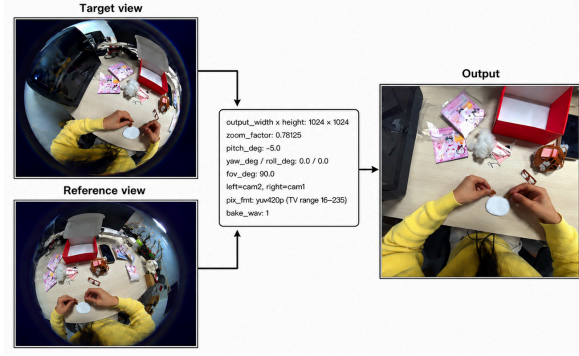


Figure 7: **Representative view-rectification example.** Raw fisheye observations are transformed into the canonical rectified egocentric view used for annotation, benchmark construction, and model training.

we use the rectified front-left RGB stream as the canonical egocentric view. The resulting videos are 1024×1024 , 20 FPS, single-view egocentric streams synchronized with audio.

The original fisheye stream provides a wide field of view but exhibits strong lens distortion. We therefore rectify the canonical front-left stream and adjust its zoom and pitch to produce a natural first-person view while retaining the hand-tool-object interaction region. A representative configuration uses a zoom factor of 0.78125, a 90° target field of view, a pitch adjustment of -5° , and zero yaw and roll. These parameters represent view-processing settings rather than semantic annotations; their purpose is to preserve the interaction region while reducing fisheye distortion.

B.2 Dense Caption Annotation

Dense captions provide temporally localized descriptions of visible actions, objects, and task progress. Each video is divided into 5-minute clips, and 32 uniformly sampled frames from each clip are first used to generate a global-context description. The clip is then processed sequentially in 5-second chunks: the first chunk is conditioned on the global context, while subsequent chunks additionally receive the preceding caption as temporal memory. Chunk-level captions are further aggregated into 1-minute windows and 5-minute summaries, providing descriptions at multiple temporal scales from fine-grained manipulation to longer task context. Object recognition on selected keyframes provides additional visual evidence during post-hoc caption checking.

B.3 Narration Annotation Guidelines

Tool-centric narrations complement surface-level descriptions by capturing the participant’s tool choice, intended use, interaction target, and the effect of the action on the task state. Contributors are instructed to narrate meaningful tool-use events rather than every low-level hand motion. A reference template encourages explicit reasoning about tool selection and feasible alternatives:

“I am using [the] [property] [tool] [location/identity cue] rather than [the] [property] [alternative tool] [location/identity cue] to [action] on/for [target] because [reason].”

The template serves only as guidance, and annotators may use natural language whenever the same information can be expressed more clearly. Representative cases include selecting narrow chopsticks rather than a spoon to reach material inside a constrained container, choosing a flat spatula rather than chopsticks to support and lift a soft object, and using a pipette rather than a larger container when finer control over a small liquid quantity is required. These examples illustrate the intended emphasis on the relationship among tool properties, task requirements, target objects, and alternative tools.

Narrations may initially be provided in English, Korean, or Chinese. Korean and Chinese materials are translated and normalized into English by bilingual project members before benchmark construction or verification, and all finalized benchmark questions and answer choices are written in English.

For selected narrations used in spatial processing, contributors annotate 2D grounding points on corresponding keyframes for focal tools or target objects selected for tracking. These points link textual mentions to visible instances and provide initialization cues for long-horizon tracking. Grounding is therefore an additional signal for selected tool-use events rather than an annotation assumed to be available uniformly for every narration or video segment.

C Benchmark Construction Details

C.1 Benchmark-Reserved Source Pool

Before instruction-data or benchmark construction, we partition the collected source videos into a training pool and a benchmark-reserved pool at the **source-video level**. The benchmark-reserved pool contains 40.34 hours of video and is selected with

consideration for annotation quality, tool-use density, task diversity, and environmental coverage. This partition is performed before constructing any training examples. Every instruction-tuning example traces to a training-pool source video, and every benchmark item traces to the reserved pool.

Within the benchmark-reserved pool, we prioritize clips containing observable hand–tool–object interactions, object-state changes, tool selection or substitution, and temporally extended procedures. Each benchmark item is associated with a localized temporal window around an anchor moment. Most clips last at least three minutes to preserve sufficient context for procedural and causal reasoning. We limit excessive temporal overlap between clips from the same source video and control QA density across source videos to reduce redundancy and maintain diversity over videos, tools, and task contexts.

C.2 Human-Crafted QA Construction

Human annotators construct 8-way multiple-choice questions with one correct answer and seven distractors. Questions are written by both original data collectors and external annotators. Original collectors contribute task-aware questions that require understanding procedural intent, tool choice, and causal effects, while external annotators identify additional visually inferable but model-challenging reasoning points. Annotators formulate questions whose answers are supported by visible evidence in the provided video. Target reasoning includes tool selection and substitution, object-state changes, contact and manipulation events, temporal ordering, observable task progress, workspace staging, and spatial hand–tool–object relations.

Domain-specific clips are assigned to annotators with relevant familiarity whenever possible. In particular, specialized procedures such as laboratory experiments are reviewed by contributors familiar with the corresponding task domain. Distractors are designed to remain plausible in the general task context while being incorrect under the observed video evidence. This process contributes human-crafted questions to the full **1,000-question benchmark**.

Language and review. All finalized EgoTools-Bench questions and answer choices are written in English. Auxiliary narrations may originate in English, Korean, or Chinese. Korean and Chinese materials are translated and normalized into En-

glish by bilingual project members before use in QA construction or verification. Final QA items are reviewed for linguistic clarity, visual answerability, and the existence of a single best answer.

C.3 Spatially Grounded QA Construction

The **1,000-question benchmark** also includes spatially grounded QA pairs generated from selected 3D annotations. The generation pipeline uses reconstructed scene geometry, focal-tool trajectories, and object-state signals to propose questions about spatial motion, placement, state changes, and tool–object interaction dynamics. Human verification of the video evidence, question validity, correct answer, and distractor quality precedes candidate admission to the benchmark. The finalized benchmark contains **1,000 8-way multiple-choice questions**.

C.4 Benchmark Evaluation Protocol

EgoTools-Bench evaluates models on questions, answers, media, and annotations derived from the benchmark-reserved source pool. During evaluation, models receive the benchmark video clip together with the question and answer choices according to the specified model protocol. Benchmark construction and analysis use corrected narrations, narration audio, dense captions, grounding annotations, source-video metadata, and 3D annotations.

D Quality Control and Normalization

D.1 Fix-First Quality Control

Raw human annotations are processed under a fix-first quality-control policy that preserves valid annotator intent whenever a defect can be repaired without changing the underlying reasoning target. Each item is evaluated using a 22-flag structural and semantic audit covering missing or malformed fields, placeholders, duplicate entries, first-person leakage, personally identifiable information, answer-option mismatch, inappropriate full-video references, temporal misalignment, inconsistent tool identity, and distractor anomalies. Content-empty submissions and verbatim duplicates are removed, whereas repairable issues are corrected using deterministic rules whenever possible. Semantically uncertain or low-confidence cases are routed to human review.

For a limited class of surface-form corrections, we use Gemini-2.5-Flash-Lite with temperature 0. The model is instructed to perform the small-

est meaning-preserving modification possible without altering the reasoning target or correct answer. Such edits include replacing identifying names with role descriptions, converting first-person references into third-person form, correcting surface grammar, and completing truncated answers only when the intended meaning is unambiguous. Model-assisted editing therefore serves as a language-normalization step rather than as a source of benchmark answers.

D.2 Eight-Choice Normalization and Anti-Shortcut Checks

All benchmark items are normalized to exactly eight answer choices. Items with insufficient distractors are augmented, items already containing seven distractors are validated, and answer sets containing too many distractors are reduced to seven plausible alternatives while preserving the original question and ground-truth answer whenever possible.

After normalization, each answer set is screened for potential textual and structural shortcuts. These checks cover meta-options such as “all of the above” and “none of the above”, duplicate or near-duplicate choices, token-set permutations, substring or superstring containment, excessive question–answer lexical overlap, answer-position imbalance, systematic correct-answer length bias, and questions that remain readily solvable without access to the video. Items flagged by deterministic, text-only, or multi-model sanity checks are manually reviewed and, when appropriate, repaired by replacing distractors with visually plausible and length-balanced alternatives. Final retention decisions depend on visual grounding, unique answerability, and annotation quality rather than on whether a particular model answers an item correctly.

To quantify answer-length bias, let L_{ans} denote the character length of the correct answer, and let μ_{dist} and σ_{dist} denote the mean and standard deviation of the seven distractor lengths. We compute

$$z = \frac{L_{\text{ans}} - \mu_{\text{dist}}}{\sigma_{\text{dist}}}.$$

Items outside the accepted $\pm 1.5\sigma$ range are flagged for repair. After applying the same normalization and length checks to all benchmark items, the finalized set has $\bar{z} \approx -0.15\sigma$, with no item outside the accepted range. Finally, answer-choice order is shuffled using a deterministic seed derived from

the QA identifier, reducing positional priors while preserving reproducibility.

D.3 Independent Annotation Reliability Study

To quantify benchmark reliability beyond the original annotation and review pipeline, we conducted an independent study on a stratified sample of **250 benchmark questions**, including **100 Affordance & Causality questions**. We deliberately included substantial coverage of Affordance & Causality because intent- and causality-oriented items are among those most likely to admit subjective interpretations.

Five independent annotators who had not written the corresponding benchmark questions participated in the study, with each annotator evaluating 100 assigned questions. This protocol yielded **500 independent judgments** in total and two judgments for every sampled item. Annotators viewed only the video, question, and eight answer choices, without access to the original answer key. In addition to selecting an answer, they assessed whether each item had a uniquely best answer supported by visible evidence.

Across all 500 judgments, independent annotators agreed with the original benchmark key in **89.8%** of cases, with **86.5%** agreement on the deliberately oversampled Affordance & Causality subset. The two independent annotators assigned to each item selected the same answer on **88.4%** of sampled questions, and **94.4%** of sampled items were judged to have a uniquely best answer supported by the video. The remaining **14 items (5.6%)** were flagged as ambiguous or having multiple plausible answers and were manually reviewed; items failing the unique-answer criterion were revised or replaced before finalization. All benchmark results reported in this paper were re-computed on the resulting quality-controlled benchmark.

E Additional Track-Wise Analysis

Precise visual grounding creates the largest cross-track separation. Perception & Grounding (PG) requires concrete visual evidence about the focal tool, target object, contact point, object attribute, quantity, or local state change. Frame-level evidence distinguishes visually similar tools, contact targets, and local state changes. Among open-source instruct models, PG accuracy ranges from 35.6% to 45.9%. Qwen3-VL-8B-Instruct,

Qwen2.5-Omni-7B-Instruct, and InternVL3.5-8B-Instruct each reach 45.9%, while LLaVA-Video-7B-Qwen2 reaches 39.7%. Gemini-3.1-Pro reaches 51.7% on PG, compared with 69.7% on Affordance & Causality, 72.1% on Procedural Dynamics, and 74.9% on Spatial Reasoning. This pattern makes PG the model’s lowest-scoring track and positions precise visual grounding as the largest cross-track opportunity for improvement.

Higher-level tracks distinguish how models use context. The Affordance & Causality (AC), Procedural Dynamics (PD), and Spatial Reasoning (SR) tracks admit different sources of contextual support, including tool-function priors for AC, task-order context for PD, and workspace geometry and physical commonsense for SR. Qwen3-VL-8B-Instruct improves over its 4B counterpart by 4.4 points on AC, 0.5 points on PG, 6.9 points on PD, and 5.4 points on SR, suggesting that additional capacity benefits procedural and spatial reasoning more strongly than precise grounding. Qwen2.5-Omni-7B-Instruct improves over Qwen2.5-VL-7B-Instruct by 4.6 points on AC, 2.6 points on PG, and 1.6 points on PD, with a 1.1-point decrease on SR. Similarly, InternVL3.5-8B-Instruct improves over InternVL3-8B-Instruct across all four tracks, with gains of 1.1 points on AC, 1.0 on PG, 1.6 on PD, and 4.4 on SR. These track-wise differences illustrate why EgoTools-Bench is diagnostic: model scale and modality produce ability-specific gains.

Model families exhibit distinct performance profiles. The model-class comparison further shows that scale, modality, and thinking-style inference produce distinct track-wise profiles. Proprietary Gemini models achieve the strongest overall results, ranging from 55.4% to 66.9%, with PG as their weakest track. Thinking-style inference is similarly heterogeneous: Qwen3-VL-8B-Thinking underperforms Qwen3-VL-8B-Instruct on AC, PG, and PD and matches it on SR, while GLM4.1V-Thinking reaches 50.7%, the strongest open-source zero-shot overall result. MiMo-VL-7B-RL reaches 45.3% overall and shows a different track profile, including 45.9% on PG and 39.1% on SR. Overall, these results show that model designs redistribute errors across the tool-mediated visual reasoning abilities.

E.1 Input-Evidence Ablation

Visual evidence and multi-frame context matter. Ordered 64-frame input achieves 51.7% overall

accuracy, compared with 42.7% for text-only input. Perception & Grounding shows the largest visual-input gain, improving by 13.9 points over text-only input. Using the middle frame gives 46.9% overall accuracy, indicating that many questions require evidence distributed across multiple moments of a tool-use process.

Temporal order provides additional information. Shuffling the same 64 frames decreases overall accuracy from 51.7% to 49.3%. Procedural Dynamics shows the largest ordered-versus-shuffled difference at 3.2 points, consistent with its emphasis on step transitions, action ordering, and changes in tool use over time. Spatial Reasoning shows a much larger ordered-versus-middle-frame gap than ordered-versus-shuffled gap, indicating stronger dependence on multi-frame coverage than on exact frame order. All conditions use the same deterministic frame-list pipeline, with the ordered condition as the matched control. Its 51.7% score belongs to this frame-list protocol; the main benchmark result uses the MP4-based protocol.

E.2 Implicit and Explicit Tool-Use Questions

Explicit tool-use cues typically improve performance. Table 6 separates questions that explicitly mention the relevant tool-use cue from questions where the model must infer it from the video context. Proprietary Gemini models show a consistent advantage on explicit questions, with explicit-minus-implicit gaps of 4.9–7.7 points. Qwen3-VL-8B-Instruct shows the same direction with a smaller 4.0-point gap. This suggests that current models benefit when the question language makes the relevant tool-use relation easier to locate. Qwen3-VL-8B-Thinking scores 45.8% on implicit and 45.2% on explicit questions, a -0.6 -point difference. Implicit questions therefore test recovery of the tool, target, and state change from video context beyond explicit wording cues.

Table 6: **Ablation on implicit and explicit tool-use questions.** Δ denotes Explicit–Implicit.

Model	Full Bench.	Implicit	Explicit	Δ
Gemini-3.1-Pro	66.9	65.2	72.9	+7.7
Gemini-3-Flash	64.4	63.1	68.0	+4.9
Gemini-3.1-Flash-Lite	55.4	54.3	59.5	+5.2
Qwen3-VL-8B-Instruct	50.0	49.1	53.1	+4.0
Qwen3-VL-8B-Thinking	45.7	45.8	45.2	-0.6

F Training and Evaluation Details

F.1 Single-Stage Training Procedure

We initialize EgoTools-8B from Qwen3-VL-8B-Instruct and perform a single stage of supervised fine-tuning using instruction examples derived exclusively from the EgoTools-Data training pool. We optimize all parameters of the language-model component while keeping the visual encoder and multimodal aligner frozen. Thus, “full-parameter” fine-tuning refers specifically to optimization of the complete language-model component rather than every module of the multimodal architecture.

We use AdamW with a constant learning rate of 2.3×10^{-6} and no warmup for one epoch. Training uses a per-device batch size of 1, gradient accumulation over 16 steps, and 8 GPUs, corresponding to an effective batch size of 128 and 1,443 optimizer steps. We use bfloat16 precision, DeepSpeed ZeRO Stage 3, FlashAttention, gradient checkpointing, a maximum multimodal sequence length of 8,192 tokens, and at most 1,024 visual tokens per image. Each training video example is represented using 64 uniformly sampled frames at a target rate of 2 FPS, with the per-video frame count fixed to 64.

Training and evaluation resources are separated before instruction-data or benchmark construction at the source-video level. No video, clip, caption, narration, synthetic QA pair, grounding annotation, spatial annotation, or other derivative of a benchmark-reserved source video is used during model training. This partition ensures that the reported EgoTools-Bench results are evaluated on source videos excluded from instruction-data construction.

F.2 Training Data Composition

The single-stage training mixture contains **184,679 EgoTools-derived instruction examples**, summarized in Table 7. Direct supervision of the four EgoTools-Bench capabilities contributes 69,760 examples (37.8% of the mixture), while the remainder provides episode-level reasoning, general video

QA, descriptive supervision, and open-ended answering. By answer format, the mixture contains 170,595 multiple-choice examples, 5,000 short open-ended examples, and 9,084 dense-captioning or narration-completion examples. Within the 5,000 next-action multiple-choice examples, answer positions are balanced uniformly across the four choices, with 1,250 examples assigned to each position.

F.3 Evaluation on Existing Egocentric Benchmarks

Existing egocentric benchmarks are evaluated using their corresponding protocols. EgoSchema is evaluated on the full 5,031-question set, EgoPlan-Bench on its 3,343-question validation set, and EgoThink as the macro-average over its twelve single-image subtasks using gpt-4o-2024-11-20 as the judge. No videos or annotations from EgoSchema, EgoPlan-Bench, or EgoThink are used during EgoTools-8B training.

As reported in Table 2, the single-stage model improves EgoTools-Bench accuracy from 50.0% to 60.9%. The gains cover Affordance & Causality (+9.3 points), Perception & Grounding (+13.8), and Procedural Dynamics (+11.0), while Spatial Reasoning decreases by 11.0 points. Spatial reasoning therefore remains a limitation of the current training recipe.

The training mixture also contains 5,000 next-action multiple-choice examples that pair a video with the current observation frame and 5,000 single-image open-ended examples. These formats mirror how EgoPlan-Bench and EgoThink pose their questions. On EgoThink, EgoTools-8B reaches 64.0% macro accuracy compared with 61.0% for the backbone, whereas on EgoPlan-Bench it reaches 35.2%, below the 42.3% backbone. Because these training subsets mirror the corresponding answer formats, we do not interpret these results as evidence of general transfer. Our primary claim therefore concerns the training utility of EgoTools-Data for tool-centric reasoning rather than uniform improvement across egocentric benchmarks.

G Dataset Statistics and Release

G.1 Dataset and Benchmark Statistics

The curated EgoTools-Data corpus contains **646 egocentric videos** totaling **100.36 hours** across seven tool-use domains: kitchen, classroom, research laboratory, repair workshop, craft, office,

Table 7: **Composition of the EgoTools-derived instruction-tuning corpus.** Counts describe the 184,679-example single-stage mixture. Shares are rounded to one decimal.

Supervised capability	# Examples	Share
Affordance & Causality (AC): tool-use purpose, intent, and causal rationale	14,175	7.7%
Perception & Grounding (PG): tool and state identification, attributes, and frame-level visual evidence	14,755	8.0%
Procedural Dynamics (PD): temporal order, workflow, and state change within a clip	22,943	12.4%
Spatial Reasoning (SR): spatial relations, placement, and hand-tool-object geometry	17,887	9.7%
Episode-level multiple choice: action ordering, activity summarization, and goal inference over a whole episode (64,554), plus next-action questions paired with the current observation frame (5,000)	69,554	37.7%
General video QA: mixed episode-level questions not targeted at a single ability	31,281	16.9%
Dense captioning and narration completion	9,084	4.9%
Single-image open-ended QA	5,000	2.7%
Total	184,679	100.0%

Table 8: **Domain-level composition of EgoTools-Data and EgoTools-Bench.**

Domain	EgoTools-Data				EgoTools-Bench
	# Videos	Hours	# Captions	# Narrations	# QAs
Kitchen	276	47.89	116,532	2,113	837
Classroom	74	4.03	24,696	467	15
Research Lab	61	37.04	111,312	1,887	99
Repair Workshop	84	3.48	44,424	763	16
Craft	67	2.04	36,612	657	23
Office	30	3.54	10,188	213	6
Household	54	2.34	17,568	419	4
Total	646	100.36	361,332	6,519	1,000

and household. It contains **361,332 hierarchical captions** and **6,519 tool-centric narrations**, together with synchronized audio where applicable and selected grounding and spatial annotations.

EgoTools-Bench is constructed from a 40.34-hour benchmark-reserved source pool and contains 1,000 8-way multiple-choice questions spanning human-crafted questions and spatial questions generated from selected 3D annotations and subsequently human-verified. Its four reasoning tracks comprise 363 Affordance & Causality questions, 236 Perception & Grounding questions, 222 Procedural Dynamics questions, and 179 Spatial Reasoning questions.

The benchmark reflects the available tool-use recordings rather than enforcing equal domain balance. As shown in Table 8, its domain distribution is strongly long-tailed, with Kitchen and Research Lab forming the largest subsets. These statistics are provided to support interpretation under the underlying domain imbalance; EgoTools-Bench is not intended to provide a domain-balanced estimate of model performance.

G.2 Release and Privacy Considerations

The public release is organized around the EgoTools-Data corpus and the EgoTools-Bench

evaluation package. Subject to participant consent and privacy review, EgoTools-Data includes rectified egocentric video clips, dense hierarchical captions, corrected English tool-centric narrations, temporal metadata, official source-video split files, selected finalized grounding annotations, and selected spatial or trajectory annotations and visualizations when cleared for release. Original narration audio is distributed only for recordings covered by the corresponding release authorization and privacy review, while raw fisheye recordings are excluded from the public release.

The EgoTools-Bench package contains the finalized QA data, answer keys, reasoning-track and domain labels, clip metadata, and the official evaluation split. To support reproducibility, the release additionally documents prompt templates, frame-sampling settings, model-specific inference configurations, invalid-answer handling rules, scoring scripts, data-loading utilities, and representative execution examples. Official split metadata preserve the source-video partition used for instruction tuning and benchmark construction so that future work can maintain the same strict evaluation separation.

All release candidates undergo privacy review for participant and bystander information as well as sensitive visual or audio content. Materials outside the applicable consent or release scope are removed or redacted, and personally identifying metadata are excluded. EgoTools is released for research use, with privacy-sensitive components governed by the applicable release authorization, anonymization requirements, and privacy-review outcomes.

H Ethics

Participation in egocentric video recording, narration, and annotation was voluntary. Participants received no monetary compensation for contributing recordings, narrations, or annotations. Before participating, contributors were informed of the project purpose, the types of data being collected, the intended research uses, and the possibility that cleared materials could be publicly released for research purposes. Participants provided consent covering the applicable collection, annotation, research-use, and release activities. No formal IRB or institutional ethics approval was sought for the EgoTools data collection; we therefore explicitly report the absence of formal institutional review and document the safeguards applied to consent, privacy, anonymization, and release.

Depending on participant role and recording setup, internally collected materials may include egocentric video, synchronized audio, spoken narrations, corrected English text narrations, benchmark QA annotations, clip metadata, timestamps, focal-tool grounding, and selected trajectory or object-state annotations. Public distribution is restricted to materials covered by the applicable release authorization and privacy review. The public package focuses on rectified video, dense captions, corrected English narrations, benchmark QA pairs, metadata, and selected finalized grounding or spatial annotations. Original narration audio is released only when the corresponding authorization and privacy checks permit it, and raw fish-eye recordings are not publicly distributed. Released metadata exclude names, contact information, institution-specific identifiers, public profiles, and other direct identifiers; sensitive or personally identifiable content is filtered, redacted, or withheld where necessary.

EgoTools-Bench is intended for academic research on tool-use reasoning, egocentric video understanding, grounded video-language modeling, and related multimodal learning problems. It is not intended to identify individuals, infer private attributes, monitor workers, or evaluate individual human ability. Dataset documentation communicates these intended uses and discourages surveillance, personally sensitive behavioral profiling, and other applications that conflict with the release scope. Contributors may request exclusion or removal of data associated with their participation in accordance with the release policy.

I Additional Related Work

Broader egocentric video reasoning. Egocentric video benchmarks have expanded from action and object understanding toward memory, planning, temporal localization, and assistance. QAEgo4D and GroundVQA emphasize episodic-memory QA and grounded temporal QA in long egocentric videos (Bärmann and Waibel, 2022; Di and Xie, 2024). EgoTaskQA evaluates task-oriented questions about dependencies, effects, goals, and beliefs (Jia et al., 2022); EgoSchema and EgoTempo focus on long-context and temporal reasoning (Mangalam et al., 2023; Plizzari et al., 2025); and EgoPlan-Bench evaluates planning from first-person videos (Chen et al., 2026). Recent benchmarks further study first-person thinking, life-log assistance, real-time aid, expert-guided assistance, situated awareness, multi-agent egocentric QA, intent prediction, personalized grounding, and AR/life-logging environments (Cheng et al., 2024; Gao et al., 2026a; Kim et al., 2026; Li et al., 2026a; Pan et al., 2026; Ragusa et al., 2026; Tang et al., 2026; Xiao et al., 2026; Yang et al., 2025; Zhou et al., 2025). These resources broaden evaluation around memory, temporal context, planning, and user assistance. EgoTools-Bench adds a complementary focus on physical tool suitability, substitution, and adaptation to changing target-object states.

Instructional, embodied, and robotic interaction benchmarks. Instructional-video datasets provide useful comparisons because they often contain tools, ingredients, techniques, and procedural alternatives. VidDetours retrieves detours from how-to videos, and StepDiff compares differences between instructional clips (Ashutosh et al., 2024; Nagarajan and Torresani, 2024); their primary scope is procedural comparison or retrieval. Embodied interaction benchmarks provide another related setting. OpenEQA evaluates open-vocabulary embodied question answering in real-world environments (Majumdar et al., 2024), RoboVQA provides large-scale video-language supervision for robotic demonstrations (Sermanet et al., 2024), and RoboCasa365 studies simulated kitchen manipulation across many tasks and environments (Nasiriany et al., 2026). EgoTools combines natural egocentric human video, tool-centric narration, narration-linked tool grounding, and explicit QA over tool choice, substitution, temporal manipulation, and state-dependent adaptation.

Affordance grounding and physical tool understanding. Robotics and embodied AI have long treated tool use as a problem of affordance, grasping, motion, and task execution. Task-oriented grasping studies how tools should be grasped for tasks such as sweeping and hammering (Fang et al., 2018), while ToolFlowNet predicts dense tool motion for manipulation tasks such as scooping and pouring (Seita et al., 2023). Broader surveys and manipulation methods study robot tool use, affordance-based manipulation, precise affordance masks, and physically grounded policies (Zitkovich et al., 2023; Gao et al., 2024; Li et al., 2025b; Qin et al., 2023; Yamanobe et al., 2017). Recent multimodal benchmarks and methods further test whether MLLMs and VLMs can recognize affordances, infer physical constraints, and ground tool-related interactions (Huang et al., 2024; Qian et al., 2024; Yu et al., 2025; Zhang et al., 2025b; Yao et al., 2026). These works measure tool functions, constraints, and multi-step interactions primarily through static images, synthetic or 3D scenes, short interactions, or constrained robotic setups. EgoTools-Bench complements them with natural first-person videos of real human tool selection, substitution, and adaptation over time.

Wearable assistants. Wearable and egocentric assistant benchmarks require situated answers from the user’s viewpoint. WearVQA studies visual question answering for wearable devices (Chang et al., 2025), while visual-intention grounding benchmarks infer user needs and intentions from egocentric observations (Sun et al., 2025); EgoTools-Bench complements them by testing tool function, substitution, ordered manipulation, and state-dependent adaptation.

J Data Acknowledgments

We thank Feiyu Li, Jialin Wu, Yichi Zhang, Wenhui Wang, Zhaoxi Chen, Jewoo Park, Jeongwon Yoo, Yulgyeom Kim, Yuan Fei, Haiheng Liu, Gordon Chen, Jiarui Zheng, Hongmei Guo and Junhan Zhu for supporting egocentric video data collection and related data preparation. We also thank all participants, annotators, reviewers, and infrastructure contributors whose efforts supported the collection, annotation, quality control, and release preparation of EgoTools.