EgoTools: Tool-Centric Reasoning

in Real-World Egocentric Videos

Shulin Tian[1,2,*] **Junsu Kim**[1,3,*] **Shuai Liu**[1,*] **Hao Li**[1,*,★] **Yujiao Shen**[1] **Sihan Li**[1] **Zhe Yang**[1] **Yeongon Kim**[1] **Feiyu Li**[4] **Jialin Wu**[5] **Yichi Zhang**[1] **Wenhui Wang**[6] **Runmao Yao**[1] **Yuhao Dong**[1] **Zhaoxi Chen**[1,★] **Fangzhou Hong**[1,★] **Antonino Furnari**[7] **Jingkang Yang**[1] **Hongyuan Zhu**[2] **Ziwei Liu**[1,†,★]

[1]S-Lab, Nanyang Technological University [2]A*STAR [3]KAIST [4]PKU [5]FDU
[6]School of Biological Sciences, Nanyang Technological University [7]University of Catania

Project Page: shulin16.github.io/egotools_dev **Code:** github.com/Ropedia/project
Dataset: huggingface.co/datasets/Ropedia/dataset **Model:** huggingface.co/Ropedia/model

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

Abstract

Real-world embodied tasks, from everyday activities to professional procedures, require agents to act under physical constraints while tracking evolving object and task states. Tool use sits at the heart of such tasks, as many everyday and professional activities are tool-mediated. Understanding them requires reasoning about affordances, hand–tool–object geometry, procedural progress, and causal effects on target objects.

* Equal contribution. † Corresponding author. ★ Ropedia Author.

Yet despite strong performance on perception-oriented video tasks such as captioning and general video QA, current multimodal video models remain limited in this form of tool-centric embodied reasoning. Progress in this direction has been limited by the lack of real-world egocentric data and diagnostic benchmarks. To address this gap, we introduce **EgoTools**, the first comprehensive suite for egocentric tool-use understanding. It consists of two complementary components: **EgoTools-Data**, a large-scale corpus of 100 hours of tool-centric egocentric recordings with synchronized audio, dense captions, reasoning-heavy narrations, and supplementary 3D information; and **EgoTools-Bench**, a diagnostic benchmark of 1,000 QA pairs across four tracks that cover tool-use understanding from perception and geometry to procedure and causal reasoning. Experiments show that current models still struggle on EgoTools-Bench, validating the difficulty of the benchmark. Beyond evaluation, we validate **EgoTools-Data** as a training resource. On the full 1,000-question benchmark, full supervised fine-tuning improves Qwen3-VL-8B-Instruct from **50.0% to 60.9%**, under strict source-video separation. Together, these results establish EgoTools as a unified resource for both training and diagnostic evaluation of real-world egocentric tool-use understanding.

1. Introduction

Human activities in the physical world, from daily chores to professional procedures, involve not only bare-hand manipulation but also the skilled use of tools. For an embodied agent, understanding such activities demands more than simply recognizing visible actions and objects. Such understanding also requires models to reason about tool–target contact, spatial coordination, procedural progress, and the physical outcomes of actions. This tight coupling of low-level perception with high-level temporal and causal awareness makes tool use a distinct challenge for embodied reasoning. The egocentric viewpoint naturally foregrounds this challenge by centering the observations of hands, tools, and their immediate consequences in a continuously evolving task state. From this perspective, the interplay of physical form, functional intent, and procedural dynamics is directly and persistently visible, making egocentric tool use a uniquely demanding and revealing testbed for embodied reasoning.

Despite strong performance on perception-oriented video tasks, current multimodal video models [2, 3, 19, 25, 61] remain limited in egocentric tool-use reasoning. Progress in this important direction has been held back by the lack of real-world egocentric data with dense tool-use annotations and by the absence of diagnostic benchmarks that isolate this reasoning. Most existing egocentric datasets [10, 20, 21, 30, 40, 44] focus on activities, objects, or hand–object interactions. Although tools are often present in these videos, they are rarely foregrounded as the central unit of explanation. In previous benchmarks [8, 9, 11, 23, 32, 37], actions may be labeled or queried at the level of "cook eggs", while the spatula and pan that mediate the activity go unmentioned, and the tools through which actions unfold are effectively invisible in their annotations. This dual absence creates both an evaluation gap and a supervision gap: current benchmarks do not isolate tool-mediated reasoning, and existing egocentric corpora rarely provide dense training signals that connect tool choice, target-object state, manipulation context, and causal outcomes.

To address this gap, we introduce **EgoTools**, the first comprehensive suite that brings together dedicated resources for both training and evaluation of egocentric tool-use understanding. At

its core, EgoTools comprises two complementary components: **EgoTools-Data**, a large-scale real-world egocentric corpus designed to provide the rich training signals that current models lack, and **EgoTools-Bench**, a carefully constructed diagnostic benchmark that isolates the specific reasoning demands of tool-mediated activities for rigorous evaluation. EgoTools-Data comprises 100 hours of tool-centric egocentric recordings spanning diverse everyday and professional tasks, with synchronized audio, dense textual annotations ranging from surface-level captions to narrations that explicitly foreground affordances, procedural structure, and causal effects, as well as supplementary 3D information. Its annotations expose learnable signals for tool choice and substitution, object-state changes, procedural progress, and grounded hand–tool–object interactions. Together, these annotations transform passive video into structured supervision for understanding how humans conduct embodied tasks with tools in the physical world. From this corpus, we construct EgoTools-Bench, a set of 1,000 question–answer pairs that systematically probe a model's capacity for embodied tool use. Our benchmark consists of four tracks: **Affordance & Causality** evaluates goal-directed reasoning about tool use, including why a tool is chosen or switched, what it affords, and how actions establish preconditions and produce consequences; **Perception & Grounding** focuses on directly observable facts such as tool and object identities, attributes, states, and quantities; **Procedural Dynamics** probes the observable organization of tool-use processes over time, including tool–action ordering, step transitions, workspace staging, and fine-grained manipulation; and **Spatial Reasoning** evaluates egocentric geometric understanding of hand–tool–object relations, including position, alignment, containment, support, contact, and depth. These tracks provide a comprehensive evaluation of tool-use understanding from perception and geometry to procedure and causal reasoning.

We evaluate representative multimodal video models on EgoTools-Bench. On the quality-controlled full 1,000-question benchmark, Qwen3-VL-8B-Instruct reaches 50.0%, indicating that substantial room remains for egocentric tool-use reasoning. We then test whether the annotations in EgoTools-Data constitute actionable supervision rather than serving only as an intermediate source for benchmark construction. Under strict source-video separation, fine-tuning Qwen3-VL-8B-Instruct on instruction data derived from the non-benchmark portion of EgoTools-Data improves accuracy from 50.0% to 60.9%. The model improves on three of the four reasoning tracks, with Spatial Reasoning the exception. Thus, EgoTools not only exposes a substantial gap in current video-language models, but also provides supervision that can partially close this tool-centric reasoning gap.

2. Related Work

Egocentric video datasets and benchmarks. Egocentric video provides the actor's visual evidence, including hands, manipulated objects, surrounding context, and temporally ordered actions [4, 13, 29, 36]. Large-scale datasets such as EPIC-KITCHENS, Ego4D, Ego-Exo4D, Assembly101, MECCANO, and HOI4D have enabled first-person action, object, hand-object, procedural, industrial, and skilled-activity understanding [10, 20, 21, 30, 40, 44]. Recent video-language benchmarks further evaluate long-context QA, temporal grounding, planning, assistance, first-person reasoning, and long-form egocentric understanding [8, 9, 11, 23, 32, 37, 55, 60]. However, these resources are typically organized around activities, objects, events, or plans, leaving explicit tool-use reasoning under-specified. Closely related instructional-video tasks study detour retrieval and step differences involving ingredients, tools, or techniques [1, 33], but they do not systematically test whether models can justify tool choice, compare feasible substitutes, or adapt manipulation to changing object states. EgoTools-Bench addresses this gap by pairing multi-environment egocentric videos with participant-provided tool-centric narrations and

Table 1 | **Comparison with representative egocentric datasets and benchmarks.** A checkmark indicates that the feature is a central design component; ✓✗ indicates partial support. Signals denote sensor/video inputs and exclude QA text or narrations used for benchmark construction: 📹 = RGB/video, ⬛ = 3D/depth/pose, ⊕ = inertial signals, and 👀 = gaze; non-RGB signals may be available only for subsets of large datasets. Hours are reported total video hours or computed from source-reported clip counts and durations; "–" denotes not applicable or not reported. Embodied interaction benchmarks are included as scope contrast rather than directly comparable egocentric video corpora.

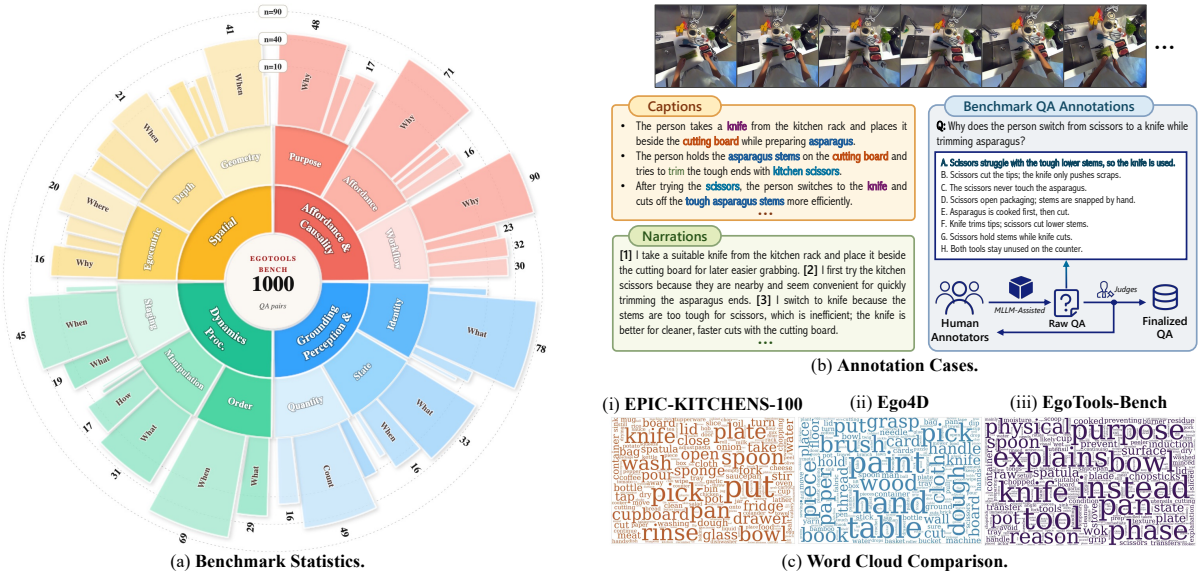
Dataset / Benchmark	Scene	Video Hours	#QA	Signals	Multi-Env.	Tool-Centric Narration	Narr.-Linked Tool Grounding	QA Benchmark	Tool Choice / Substitution	Physical Tool-Use Reasoning
— Egocentric Observation Datasets —										
EPIC-KITCHENS [10]	Kitchen	100	–	📹	✗	✗	✗	✗	✗	✗
Ego4D [20]	Real-world	3,670	14.5 K	📹⬛⊕👀	✓	✗	✗	✓✗	✗	✗
Ego-Exo4D [21]	Skilled	1,286	–	📹⬛⊕👀	✓	✗	✗	✗	✗	✓✗
Assembly101 [44]	Assembly	513	–	📹⬛	✗	✗	✗	✗	✗	✗
MECCANO [40]	Industrial	6.9	–	📹⬛👀	✗	✗	✗	✗	✗	✓✗
HOI4D [30]	Indoor HOI	44.4	–	📹⬛	✗	✗	✗	✗	✗	✓✗
— Egocentric Video Reasoning Benchmarks —										
EgoSchema [32]	Real-world	250+	5,031	📹	✓	✗	✗	✓	✗	✗
EgoTaskQA [23]	Task videos	14	40 K	📹	✗	✗	✗	✓	✗	✓✗
EgoPlan-Bench [8]	Daily tasks	–	4,939	📹	✗	✗	✗	✓	✗	✓✗
EgoIntent [35]	Daily tasks	2.9	3,014	📹	✗	✗	✗	✓	✓✗	✓✗
GroundVQA [11]	Long videos	736	303 K	📹	✗	✗	✗	✓	✗	✗
EgoTempo [37]	Temporal	6.3	500	📹	✗	✗	✗	✓	✓✗	✓✗
— Embodied Interaction Benchmarks —										
OpenEQA [31]	Indoor EQA	–	1.6 K	📹⬛	✓	✗	✗	✓	✗	✗
RoboVQA [45]	Robotics	238	829 K	📹	✓✗	✗	✗	✓	✓✗	✓✗
RoboCasa365 [34]	Sim kitchens	2,200+	–	📹⬛	✓	✗	✗	✗	✓✗	✓✗
EgoTools (Ours)	**Real-world**	100	1,000	📹⬛⊕	✓	✓	✓	✓	✓	✓

explicit grounding to the focal tool.

Physical tool-use reasoning. Physical tool use requires reasoning beyond object categories: a model must connect a tool's function and affordances to the current goal, target material, contact and motion constraints, and evolving object state. Robotics has studied related problems through tool-use surveys, task-oriented grasping, tool-flow prediction, affordance-centric manipulation, and egocentric affordance learning [12, 27, 39, 43, 54]. These works are valuable for execution, but they often focus on structured tasks, constrained manipulation, or short interaction episodes. A parallel line evaluates MLLMs and VLMs on affordance grounding and physical tool understanding [22, 38, 57, 59]; many such settings rely on static images, 3D scenes, synthetic data, isolated interactions, or robotics setups. EgoTools-Bench complements both lines with human-centered first-person evidence, where real tools are selected, substituted, and adapted within continuous tasks. Accordingly, EgoTools evaluates whether models can connect visual evidence to functional choices, feasible alternatives, temporal manipulation steps, and changes in object state.

3. EgoTools Data Suite

This section introduces EgoTools, a real-world egocentric video suite for physically grounded tool-use understanding. We first describe the collection and preprocessing of EgoTools-Data, a 100-hour real-world egocentric video corpus captured with synchronized video, audio, and geometry-related signals. We then introduce its multimodal annotation pipeline, including hierarchical dense captions, tool-centric narrations with 2D grounding, and 3D annotations derived from reconstruction and long-horizon object tracking. Finally, we describe two complementary resources derived from EgoTools-Data and separated at the source-video level: (i) an EgoTools-derived video instruction-tuning corpus constructed from the training pool, and (ii)



(a) **Benchmark Statistics.** (b) **Annotation Cases.** (c) **Word Cloud Comparison.**

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.

EgoTools-Bench, a 1,000-question 8-way multiple-choice diagnostic benchmark constructed from a curated 40.34-hour benchmark-reserved pool. EgoTools-Bench comprises 900 human-crafted questions and 100 human-verified spatial questions across four tracks: Affordance & Causality, Perception & Grounding, Procedural Dynamics, and Spatial Reasoning.

3.1. Data Collection and Preprocessing

We collect raw egocentric videos with **HOMIE**[1], a lightweight head-mounted multimodal recording device designed with four synchronized fisheye camera views, together with audio and auxiliary motion/synchronization signals that support downstream spatial processing. These signals support 3D processing, including reconstruction, pose/depth estimation, and object/hand tracking. For annotation and training, we use the rectified front-left RGB stream as the canonical view, applying zoom and pitch adjustments to obtain a natural first-person perspective while preserving hand–tool–object interactions. Rectification details and examples are provided in Appendix D.1. The final videos are 1024×1024, 20 FPS, single-view egocentric streams synchronized with audio.

Our data collection is organized around two broad contexts: **daily activities** and **expertise-intensive procedures**. Within these contexts, we collect videos across seven tool-use domains: kitchen, classroom, research lab, repair workshop, craft, office, and household, shown in Figure 1. These domains cover diverse tool-mediated activities, from everyday manipulation to craft, scientific, and fabrication procedures. Household recordings capture everyday tool use, craft and repair-workshop recordings emphasize material transformation and manual operations, and laboratory recordings include both educational and professional experiments requiring specialized knowledge and expert tool handling. Data collection was conducted across multiple kitchens, workshops, laboratories, and daily-living spaces at universities and research sites in

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

Asia. Our data collectors include graduate and undergraduate students from diverse disciplinary backgrounds, with domain expertise matched to the task whenever specialized knowledge is required. In particular, expertise-intensive recordings are performed or reviewed by collectors familiar with the corresponding procedures, ensuring that the captured tool use is both natural and technically valid. Details of the collection domains and anonymized participant information are provided in Appendix C.

The collected and curated EgoTools-Data corpus contains approximately **100 hours** of egocentric video across the seven tool-use domains described above. To balance controlled task coverage with natural tool-use behavior, we adopt two complementary collection settings: **structured task-guided** and **open-ended participant-driven**. In the structured task-guided setting, participants follow predefined task sequences prepared by the data collection team. Each sequence specifies a task goal and key procedural steps, yielding clear task boundaries, observable task-state changes, and controlled coverage of hand–tool–object interaction. In the open-ended participant-driven setting, participants are given only a broad topic or high-level goal and complete the task in their own manner. This setting allows spontaneous tool choices, procedural adaptation, repeated attempts, error recovery, and opportunistic substitutions. Together, the structured setting provides consistent procedural data for reliable annotation and benchmark construction, while the participant-driven setting captures the variability and adaptivity of in-the-wild tool use.

Source-video partition. Before constructing the instruction-tuning corpus and benchmark, we partition the recordings at the source-video level into a training pool and a benchmark-reserved pool. The 40.34-hour benchmark pool is held out from all stages of instruction-data construction. Consequently, no clip, caption, narration, synthetic QA, or other annotation derived from a benchmark source video is included in model training.

3.2. Multimodal Annotation and Curation

Textual Annotations. We provide two complementary textual annotations: **dense captions** and **tool-centric narrations**. Dense captions describe visible actions and task progress across temporal scales, while narrations capture participant intent, tool grounding, and tool-use reasoning. For dense captioning, we use Gemini-3-Flash [17] in a streaming hierarchical pipeline: each video is split into 5-minute clips, with 32 uniformly sampled frames used to generate a global context caption. Each clip is then captioned sequentially in 5-second chunks, where the first chunk uses the global caption and later chunks use the previous caption as memory for temporal consistency. Chunk captions are further aggregated into 1-minute windows and 5-minute summaries. In parallel, object recognition on key frames provides visual evidence for post-hoc hallucination checks. Following common practice in egocentric video benchmarks, where narrations serve as weak supervision for first-person activities [10, 20], we also collect narrations for EgoTools. Unlike prior datasets that encourage broad descriptions of visible actions, our narrations are explicitly tool-centric: collectors who perform the tasks narrate meaningful tool-use events using a reference narration template, emphasizing tool selection, usage intent, and effects on target objects or task states. For selected narrations, collectors annotate 2D grounding points on corresponding keyframes for focal tools or objects chosen for tracking, linking textual mentions to visual instances and providing initialization cues for long-horizon 3D tracking. The dense captions and corrected English narrations are subsequently converted into temporally grounded video instruction examples, providing supervision for action description, procedural summarization, tool selection and substitution, object-state tracking, and next-step prediction. The textual annotation interfaces are shown in Figure 3.

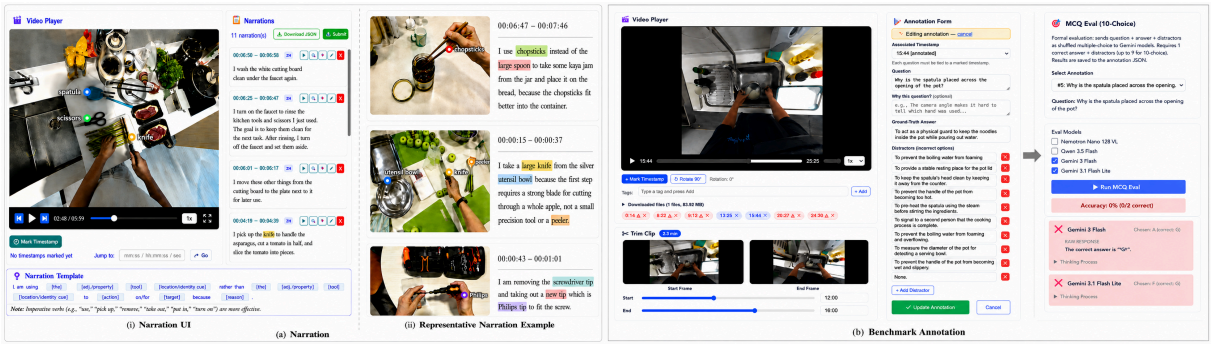


Figure 3 | **Textual annotation pipeline.** (a) Tool-centric narrations are annotated from grounded egocentric video segments. (b) Benchmark QA pairs are constructed and refined through an annotation interface with MLLM-assisted checking.

Additional details on the narration template and examples are provided in Appendix D.2.

3D Annotations. As shown in Figure 4, we collect raw geometric data with the HOMIE device from Ropedia, Inc., including egocentric videos, camera poses, and depth maps. Following Holi-Spatial [16], we train a 3D Gaussian Splatting model [28] to build a multi-view consistent scene geometry, reduce artifacts such as floaters, and render high-fidelity, temporally coherent depth maps for each frame.

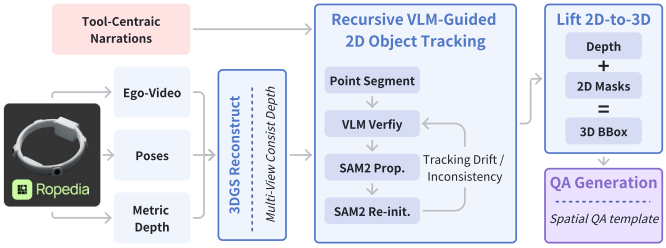


Figure 4 | **3D annotation pipeline.** Raw geometry from the HOMIE device is reconstructed with 3DGS to obtain consistent depth, then combined with tool-centric narrations for recursive VLM-guided object tracking. The resulting masks are lifted to 3D for object boxes and spatial QA generation.

For robust long-term object tracking, we introduce a recursive VLM-guided segmentation framework. Tool-centric narrations provide initial 2D grounding points and semantic captions, from which SAM2 [42] propagates object masks through the video. When later frames contain additional grounding points, a VLM [19] verifies tracking consistency. If drift or semantic mismatch is detected, the framework re-initializes SAM2 from the failure point and refines the segmentation. This feedback loop yields spatio-temporal masks with both geometric precision and long-range semantic coherence. Finally, we lift the tracked masks into 3D using the reconstructed depth and camera poses, enabling object boxes and long-horizon 4D object trajectories. Based on these annotations, we design spatial QA templates that query object motion and spatial relations. Together, the original 100-hour egocentric videos, dense textual annotations, and 3D annotations constitute EgoTools-Data, from which we derive a video instruction-tuning corpus and a source-video-disjoint EgoTools-Bench evaluation set.

3.3. Video Instruction-Tuning Data Construction

To evaluate whether EgoTools-Data provides actionable training supervision, we construct a video instruction-tuning corpus exclusively from the EgoTools training pool. Every training instance is derived from EgoTools videos and their associated dense captions, corrected English narrations, and temporal metadata. The instruction-tuning corpus contains **184,679** examples, and every example is labeled with the capability it supervises. **Grounding-oriented** supervision

asks what is visible and where: which tool is in use, which object it contacts, what attribute or local state has changed, and how hands, tools, and objects are arranged in space. **Reasoning-oriented** supervision asks why and in what order: why a tool is chosen for a material or goal, what the causal effect of an action is, how a task progresses through its steps, and what happens next. A third group provides **descriptive** supervision through dense captioning and narration completion, which teaches the model to verbalize first-person visual evidence without a question format.

Concretely, 69,760 examples (37.8%) supervise the four EgoTools-Bench abilities directly, split into affordance and causality (14,175), perception and grounding (14,755), procedural dynamics (22,943), and spatial reasoning (17,887). A further 69,554 examples (37.7%) are episode-level multiple-choice questions about action ordering, overall activity, and task goals, including 5,000 next-action questions that pair a video with the current observation frame. General video QA contributes 31,281 examples (16.9%) of mixed episode-level questions. The remainder is descriptive or open-ended: 9,084 examples (4.9%) of dense captioning and narration completion, and 5,000 (2.7%) of single-image open-ended QA. Synthetic QA candidates are filtered for visual answerability, temporal grounding, single-best-answer validity, and resistance to textual shortcuts, and answer letters are balanced within each option-count group to reduce positional bias. Figure 5 summarizes the resulting distribution.

3.4. Benchmark Construction and Quality Control

Benchmark QA Construction. Based on EgoTools-Data, we select a high-quality **40.34-hour** benchmark-reserved subset for EgoTools-Bench construction, prioritizing annotation quality, tool-use density, task diversity, and environmental coverage. The benchmark and instruction-tuning corpus are disjoint at the source-video level. We manually create QA pairs with both original data collectors and external annotators: collectors contribute task-aware procedural and causal questions, while external annotators identify visually inferable but model-challenging reasoning points. Domain-specific videos, such as wet-lab procedures, are assigned to annotators with relevant expertise. Annotators identify segments requiring physically grounded tool-use reasoning, including tool selection, object-state changes, and temporal ordering of tool-mediated actions, and formulate 8-way multiple-choice questions with one correct answer and seven distractors. Each benchmark clip is extracted as a localized temporal window around an anchor moment, with most clips lasting at least three minutes. We limit temporal overlap and control QA density across source videos to maintain diversity over videos, tools, and task contexts. This yields **900** human-crafted QA pairs. We further generate **100** spatially grounded QA pairs from 3D annotations using the Holi-Spatial [16] pipeline, augmented with focal-tool trajectories and object-state signals. These questions target spatial motion, state changes, and tool–object interaction dynamics, and are included only after human verification. Overall, the average temporal span per QA is **4.19 minutes**.

Quality Control. All benchmark items undergo a fix-first quality-control pipeline that repairs valid annotator intent before removal. We audit structural validity, video grounding, and answer-choice quality to detect malformed questions, duplicates, empty submissions, missing visual evidence, temporal misalignment, inconsistent tool identity, and weak distractors. Empty or verbatim-duplicate items are removed; repairable issues are fixed with deterministic rules or minimal model-assisted edits, including reference normalization, first-person leakage removal, and unambiguous answer completion. Low-confidence cases are sent to human review. After repair, questions are normalized to an 8-way multiple-choice format and screened for shortcuts such as near-duplicate choices, revealing wording, lexical imbalance, answer-length

or position bias, and text-only solvability. Text-only and multi-model sanity checks are used to flag potentially trivial or shortcut-solvable items for review. Flagged items are hardened by replacing distractors with visually plausible, length-balanced alternatives while preserving the original question and ground-truth answer when possible. Final revision and retention decisions are based on visual grounding, unique answerability, and annotation quality rather than the correctness of any particular model. Details are in Appendix F.

Independent Reliability Study. To quantitatively assess annotation reliability, we conducted an independent study on a stratified sample of 250 benchmark questions, including 100 Affordance & Causality questions. Five annotators who had not written the original questions each evaluated 100 assigned items, such that every question received two independent judgments. Annotators viewed only the video, question, and answer choices without access to the original answer key. Across 500 judgments, agreement with the original key was **89.8%**, including 86.5% on Affordance & Causality questions. Exact agreement between the two assigned annotators was **88.4%** at the item level, and **94.4%** of items were judged to have a uniquely best answer supported by visible evidence. Fourteen items (**5.6%**) were flagged as ambiguous or having multiple plausible answers; all were manually reviewed, and items failing the unique-answer criterion were revised or replaced. All results reported in this paper were recomputed on the resulting quality-controlled benchmark.

3.5. Distribution

Figure 2 summarizes the composition of EgoTools-Bench, which contains **1,000 QA pairs** across four tracks: **1) Affordance & Causality** (**363**), **2) Perception & Grounding** (**236**), **3) Procedural Dynamics** (**222**), and **4) Spatial Reasoning** (**179**). These tracks cover core aspects of egocentric tool-use understanding, including tool recognition, object-state grounding, hand–tool–object relations, task progress, and the causal effects of tool use. This distribution reflects the goal of EgoTools-Bench: evaluating tool use as a reasoning problem involving tool selection, application, and effects, while retaining coverage of perception, procedure, and spatial reasoning. Each track is further divided into fine-grained subtracks for detailed analysis of model strengths and failure modes. The seven collection domains characterize the coverage of the full EgoTools-Data corpus rather than a domain-balanced benchmark. EgoTools-Bench exhibits a strongly long-tailed distribution across domains, with the majority of questions drawn from Kitchen videos; complete domain-level statistics are provided in Appendix E.1. We therefore focus our primary analysis on reasoning capabilities rather than domain-level performance. To avoid over-representing individual recordings, we limit temporal overlap among benchmark clips and cap QA density per source video. Figure 2 also compares the vocabulary distribution of EgoTools-Bench with EPIC-KITCHENS-100 and Ego4D. The word clouds qualitatively illustrate that EgoTools-Bench places greater lexical emphasis on tools, manipulated objects, actions, and state changes than on general egocentric activity recognition.

4. Experiments

4.1. Experimental Setup

We evaluate EgoTools-Bench with human experts, proprietary models, and open-source models, and compare model performance with existing egocentric video benchmarks including EgoSchema [32], EgoPlan [8], and EgoThink [9]. The model set covers standard video-language models, audio-capable omni models, and thinking-style variants from Gemini [19], Qwen [2, 3,

53, 50], InternVL [51, 61], LLaVA [25, 58], MiMo-VL [48], and GLM [49] families. For EgoTools-Bench, we report overall accuracy and track-wise accuracy following the four tracks introduced in Section 3.5; existing benchmark results are reported when available under comparable settings. For each model, we report the number of input frames and whether audio is used. Most open-source instruct models use 64 uniformly sampled frames; thinking models use 64 or 512 frames depending on their supported setting; Gemini [19] models use 1 FPS video input. Audio-enabled models receive only synchronized non-narration audio. Corrected narrations, narration audio, dense captions, and annotation metadata are never provided as input.

Human evaluation. Human performance was measured using five evaluators who had not authored the questions they evaluated. Each evaluator answered all 1,000 benchmark questions over multiple sessions and could freely seek and replay the videos. Evaluators did not have access to the answer key, and there were no skipped or invalid responses. Overall and track-wise accuracies are micro-averaged over 5,000 individual judgments. The human results therefore reflect a free-viewing condition rather than the fixed-frame input budget used for model evaluation.

EgoTools training. To evaluate the training utility of EgoTools-Data, we adapt Qwen3-VL-8B-Instruct with a single stage of full-parameter supervised fine-tuning of its language-model component, while keeping the visual encoder and multimodal aligner frozen. The mixture is organized by what each example supervises: 69,760 examples (37.8%) target the four EgoTools-Bench reasoning abilities directly; 69,554 (37.7%) are episode-level multiple-choice questions about action ordering, overall activity, and task goals, including next-action prediction; 31,281 (16.9%) are general video QA; and the remainder is descriptive or open-ended, with 9,084 (4.9%) dense captioning and narration completion and 5,000 (2.7%) single-image open-ended QA. Figure 5 visualizes this distribution, and Table 10 gives the full breakdown. No videos, images, or QA annotations from EgoSchema, EgoPlan-Bench, or EgoThink are used, and all training videos are disjoint from EgoTools-Bench at the source-video level. Full training details are provided in Appendix H.

4.2. Main Results

EgoTools-Bench is challenging for current video-language models. Table 2 summarizes performance across human experts, proprietary models, open-source instruct models, and open-source thinking models. EgoTools-Bench uses an 8-way multiple-choice format, so chance performance is 12.5%. Across open-source instruct models, performance remains low, with overall accuracy ranging from 38.9% to 50.0%, far below the human expert score of 83.2%. The strongest open-source instruct result is Qwen3-VL-8B-Instruct [2] at 50.0%, followed by InternVL3.5-8B-Instruct [51] at 48.7% and Qwen2.5-Omni-7B-Instruct [53] at 47.1%. Among open-source thinking models, GLM4.1V-Thinking [49] achieves the strongest overall result at 50.7%. These results indicate that EgoTools-Bench exposes a substantial gap in current models' ability to reason about tool-mediated actions, rather than an isolated weakness of one architecture. The final row reports EgoTools-8B, our model fine-tuned on EgoTools-Data; it is listed as a separate group because, unlike every row above it, it is not evaluated zero-shot, and we analyze it in §4.3.

Existing benchmarks do not fully transfer to tool-use reasoning. Several models perform substantially better on existing egocentric benchmarks than on EgoTools-Bench. For example, Qwen3-VL-8B-Instruct reaches 69.0% on EgoSchema [32] and 61.0% on EgoThink [9], but only 50.0% on EgoTools-Bench. Qwen3-VL-4B-Instruct shows the same gap, with 69.6% on

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 our EgoTools-Bench benchmark. 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.

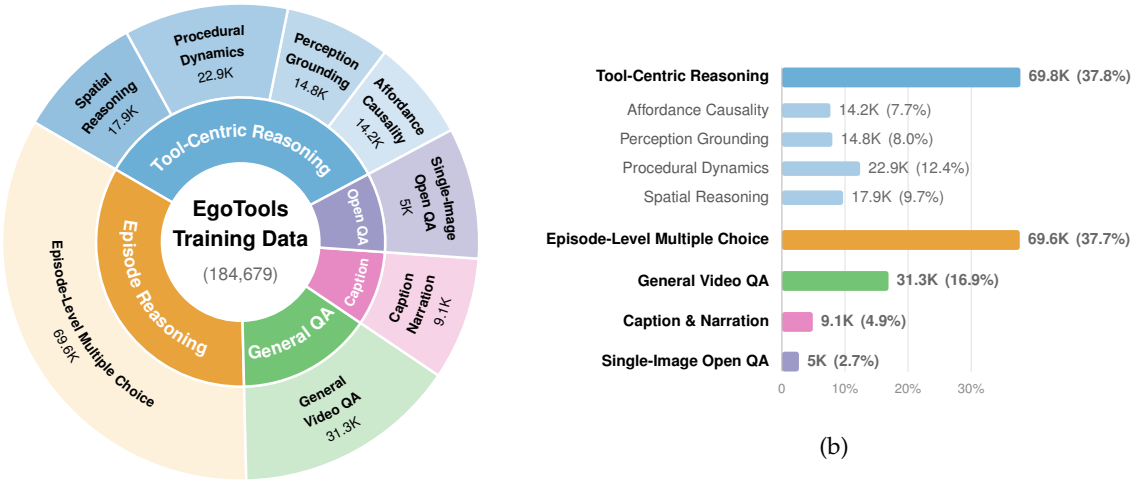
Model	Frames	Audio	EgoSchema [32]	EgoPlan [8]	EgoThink [9]	EgoTools				
						AC	PG	PD	SR	Overall
Human Baseline										
Human Expert	–	✓	~76	–	–	82.1	85.2	83.3	82.7	83.2
Proprietary Models										
Gemini-3.1-Pro [19]	1fps	✓	–	–	–	**69.7**	**51.7**	72.1	**74.9**	**66.9**
Gemini-3-Flash [17]	1fps	✓	–	–	–	64.5	50.0	**73.0**	72.6	64.4
Gemini-3.1-Flash-Lite [18]	1fps	✓	–	–	–	58.1	44.5	59.9	58.7	55.4
Open-Source Models (Instruct)										
Qwen3-VL-8B-Instruct [2]	64	✗	69.0	**42.3**	61.0	**46.1**	**45.9**	**57.1**	**55.4**	**50.0**
Qwen3-VL-4B-Instruct [2]	64	✗	**69.6**	42.0	**61.9**	41.7	45.4	50.2	50.0	45.7
Qwen2.5-VL-7B-Instruct [3]	64	✗	61.0	32.0	57.7	40.1	43.3	50.2	47.8	44.3
Qwen2.5-Omni-7B-Instruct [53]	64	✓	65.2	–	58.9	44.7	**45.9**	51.8	46.7	47.1
Qwen2-VL-7B-Instruct [50]	64	✗	63.6	39.9	60.1	46.6	35.6	46.2	47.8	44.2
InternVL3.5-8B-Instruct [51]	64	✗	63.8	–	56.0	45.8	**45.9**	53.4	53.3	48.7
InternVL3-8B-Instruct [61]	64	✗	**69.6**	–	59.4	44.7	44.9	51.8	48.9	47.1
LLaVA-OneVision-7B [25]	64	✗	63.4	–	54.9	37.6	37.1	42.9	37.0	38.9
LLaVA-Video-7B-Qwen2 [58]	64	✗	52.4	–	58.0	38.2	39.7	43.3	37.0	39.8
MiMo-VL-7B-SFT [48]	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 [2]	512	✗	**70.3**	**43.6**	62.0	43.6	41.8	48.2	**55.4**	45.7
Qwen3-VL-4B-Thinking [2]	512	✗	59.8	43.1	62.3	37.6	36.6	36.4	41.3	37.4
MiMo-VL-7B-RL [48]	64	✗	57.0	36.6	59.4	43.1	45.9	50.6	39.1	45.3
GLM4.1V-Thinking [49]	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

EgoSchema, 61.9% on EgoThink, and 45.7% on EgoTools-Bench. Thus, broad first-person activity understanding does not necessarily imply robust tool-use reasoning about affordances, grounding, procedures, spatial relations, and state changes.

Additional model and input capacity do not close the gap. Model parameter scale, audio support, and thinking-style inference mixed rather than consistent gains in overall accuracy. The gap is therefore distributed rather than localized, and we decompose it by track and model class in §4.4, then by tool-use cues and error patterns in §4.5. These results motivate a complementary question: whether the missing tool-centric capabilities can be learned from the supervision provided by EgoTools-Data.

4.3. Training Utility of EgoTools-Data

EgoTools supervision improves tool-centric reasoning. Under the standard 64-frame MP4-based evaluation protocol, single-stage full supervised fine-tuning on 184,679 EgoTools-derived instruction examples improves Qwen3-VL-8B-Instruct from **50.0% to 60.9%** on EgoTools-Bench, a gain of 10.9 percentage points. The improvement covers three of the four tracks: +9.3 points on Affordance & Causality, +13.8 on Perception & Grounding, and +11.0 on Procedural Dynamics, while Spatial Reasoning decreases by 11.0 points. Because the training and evaluation resources are separated at the source-video level, these gains reflect generalization to held-out tool-use episodes rather than memorization of benchmark clips. The result demonstrates that EgoTools-



(a) (b)

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; segment angles are schematic rather than strictly proportional, so that the smallest categories remain readable. (b) Shares drawn to scale; the four tool-centric categories are nested under their group total. Exact numbers are listed in Table 10.

Data is not merely an intermediate source for benchmark construction, but provides actionable supervision for learning egocentric tool-use reasoning.

A rebalanced single-stage mixture removes the need for a separate refinement stage. An earlier version of this recipe trained on a broader 220,334-example mixture and lost substantial EgoSchema accuracy, which had to be recovered by a second LoRA refinement stage on multiple-choice data. Rebalancing the first-stage mixture makes that stage unnecessary: the single-stage model reaches **67.6%** on EgoSchema, within 1.4 points of the 69.0% backbone, without any additional refinement and without using any original EgoSchema videos or QA annotations.

The EgoPlan and EgoThink results are consistent with answer-format coverage. Beyond multiple-choice video QA, the training mixture covers two further answer formats built from EgoTools training videos: 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 match how EgoPlan-Bench and EgoThink pose their questions. On EgoThink, the model reaches **64.0%** macro accuracy, exceeding the 61.0% backbone by 3.0 points, whereas on EgoPlan-Bench it reaches 35.2%, still 7.1 points below the 42.3% backbone. Because these training subsets mirror the answer formats used by EgoPlan-Bench and EgoThink, we do not interpret these results as evidence of general transfer. We therefore limit our primary claim to the training utility of EgoTools-Data for tool-centric reasoning rather than claiming uniform improvement across all egocentric benchmarks.

4.4. Analysis

EgoTools training improves multiple reasoning abilities. Table 3 shows that full supervised fine-tuning improves over the original Qwen3-VL-8B-Instruct backbone. Overall accuracy increases from **50.0% to 60.9%**, with gains of 9.3 points on Affordance & Causality (AC), 13.8 points on Perception & Grounding (PG), and 11.0 points on Procedural Dynamics (PD), against a

Table 3 | **Training utility of EgoTools-Data.** EgoTools-8B is Qwen3-VL-8B-Instruct after a single stage of full supervised fine-tuning on 184,679 EgoTools-derived instruction examples. Both models use the same standard 64-frame MP4-based evaluation protocol. The instruction-tuning corpus and EgoTools-Bench are disjoint at the source-video level. EgoTools-Bench results are reported on the full quality-controlled 1,000-question benchmark; EgoSchema is reported on the full 5,031-question set and EgoThink as the macro-averaged score over its twelve subtasks.

Model	EgoTools-Bench					Existing Egocentric Benchmarks		
	AC	PG	PD	SR	Overall	EgoSchema	EgoPlan	EgoThink
Qwen3-VL-8B-Instruct	46.1	45.9	57.1	**55.4**	50.0	**69.0**	**42.3**	61.0
EgoTools-8B	**55.4**	**59.7**	**68.1**	44.4	**60.9**	67.6	35.2	**64.0**

decrease of 11.0 points on Spatial Reasoning (SR). The gains on AC, PG, and PD indicate that the benefit of EgoTools supervision is not confined to a single question category, and that the training corpus provides useful supervision for visual grounding and temporal process understanding. The SR regression may reflect the limited proportion of explicit spatial supervision in the overall training mixture.

Perception & Grounding is the diagnostic anchor. Perception & Grounding (PG) is particularly dependent on concrete visual evidence because it asks about the focal tool, target object, contact point, object attribute, quantity, or local state change. Activity context alone rarely determines which visually similar tool is being used, which object it contacts, or what has changed locally. Open-source instruct models score between 35.6% and 45.9% on PG, and even Gemini-3.1-Pro reaches 51.7%, still 18–23 points below its AC, PD, and SR scores. Together with the input ablation in Table 4, where PG shows the largest improvement from visual evidence over text-only input, these results identify precise visual grounding as a persistent bottleneck.

Model design affects reasoning tracks unevenly. The higher-level tracks admit different sources of contextual support: 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 most strongly on PD and SR, by 6.9 and 5.4 points, respectively, while improving PG by only 0.5 points. This suggests that additional capacity benefits procedural and spatial reasoning more than precise grounding. Audio and thinking-style inference are similarly inconsistent: Qwen2.5-Omni-7B-Instruct improves over Qwen2.5-VL-7B-Instruct on AC, PG, and PD but scores slightly lower on SR, while Qwen3-VL-8B-Thinking underperforms its instruct counterpart on AC, PG, and PD and only matches it on SR. In contrast, GLM4.1V-Thinking achieves the strongest open-source result in Table 2, at 50.7%. These heterogeneous patterns show that scale, modality, and deliberation change where models fail, but do not uniformly resolve the need to ground reasoning in the relevant tool-mediated evidence.

4.5. Ablations and Error Analysis

EgoTools-Bench requires visual evidence beyond language priors. Table 4 compares temporally ordered 64-frame input with the same frames in shuffled order, a single middle frame, and text-only input on the full 1,000-question benchmark. All conditions yield zero parsing or inference failures. Ordered video achieves 51.7% overall accuracy, compared with 42.7% for text-only input, corresponding to a 9.0-point improvement. The substantial gain from visual input confirms that the benchmark cannot be solved solely through answer-choice patterns, tool-function commonsense, or language priors. Text-only performance nevertheless remains above

Table 4 | **Input-evidence ablation on the full 1,000-question EgoTools-Bench.** All conditions use the same Qwen3-VL-8B-Instruct backbone, deterministic decoding, and scoring configuration. The visual conditions use the same pre-extracted frame-list pipeline, with the ordered condition serving as the matched within-protocol control.

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

the 12.5% chance level. This indicates that some questions retain exploitable commonsense or answer-option cues, despite the benchmark's anti-shortcut filtering. We therefore interpret text-only performance as evidence of residual language priors rather than as evidence that video input is unnecessary.

Multi-frame coverage is important. Using only the middle frame reduces overall accuracy from 51.7% to 46.9%, showing that a single static observation is insufficient for many questions. The 4.8-point gap indicates that models benefit from observing multiple moments of the tool-use process, including the selection of a tool, its interaction with a target object, intermediate state changes, and the resulting task outcome. The effect is particularly strong for Spatial Reasoning, where ordered multi-frame input outperforms the middle-frame condition by 11.9 points. This result suggests that spatial questions frequently require integrating evidence distributed across the clip rather than reading a single hand–tool–object configuration.

Temporal order provides additional information. Shuffling the same 64 frames reduces overall accuracy from 51.7% to 49.3%. Because the visual content and frame count remain unchanged, the 2.4-point difference isolates the contribution of temporal ordering. Procedural Dynamics is the most order-sensitive track, with ordered input outperforming shuffled input by 3.2 points. This is consistent with the track's emphasis on step transitions, action ordering, preparation, and changes in tool use over time. By contrast, Spatial Reasoning shows only a 1.1-point difference between ordered and shuffled frames. Together with its large ordered-versus-middle-frame gap, this pattern indicates that SR benefits primarily from multi-frame coverage, while depending less strongly on the precise temporal sequence of those frames.

The benchmark tracks require complementary evidence. The ablation effects differ systematically across the four tracks. Perception & Grounding shows the largest dependence on visual evidence, improving by 13.9 points over text-only input. Procedural Dynamics is the most sensitive to temporal order, while Spatial Reasoning benefits most from multi-frame coverage. These complementary patterns support the intended diagnostic roles of the benchmark tracks: PG tests grounding in concrete visual evidence, PD tests temporally ordered process understanding, and SR tests the integration of spatial evidence across multiple observations. All ablation conditions use the same deterministic frame-list pipeline. The ordered condition therefore serves only as the matched control for comparisons within this ablation. Its absolute score is not directly compared with the separate main benchmark result obtained using the standard MP4-based evaluation protocol.

Implicit tool-use questions are generally harder. Table 5 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 is the exception, with a small negative gap of 0.6 points, suggesting that longer deliberation does not necessarily make the model more responsive to explicit cue wording. Implicit questions are therefore useful as a stress test for whether a model can recover the tool, target, and state change from visual evidence rather than relying on wording cues.

Table 5 | **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

Error overlap reveals systematic failure patterns. The error map in Figure 6 compares the Jaccard overlap between models' incorrect predictions. The strongest off-diagonal overlap is between Qwen3-VL-8B-Instruct and Qwen3-VL-8B-Thinking (0.63), showing that thinking-style inference leaves many of the same failures unresolved despite a slightly different accuracy profile. Gemini models also share a moderate error profile with one another, especially Pro and Flash (0.53), but their overlap with Qwen3-VL-8B is lower (0.35–0.38 for Pro/Flash), indicating that stronger proprietary models avoid some open-source errors rather than simply improving uniformly on the same examples. Gemini-3.1-Flash-Lite sits between these regimes, with higher overlap with Qwen3-VL-8B-Instruct (0.50) and Qwen3-VL-8B-Thinking (0.46). Overall, the overlap pattern supports that failures are systematic and partially model-family-specific: thinking-style variants retain many failures of their instruct counterparts, while proprietary and open-source models exhibit more complementary error sets. Taken together, the track-wise results, implicit-explicit ablation, and error-overlap analysis show that EgoTools-Bench is not merely harder than existing egocentric benchmarks. It diagnoses whether models can ground tool-mediated interactions in visual evidence, reason about tool choice and state change, and avoid substituting language, task-order, or commonsense priors for required video evidence.

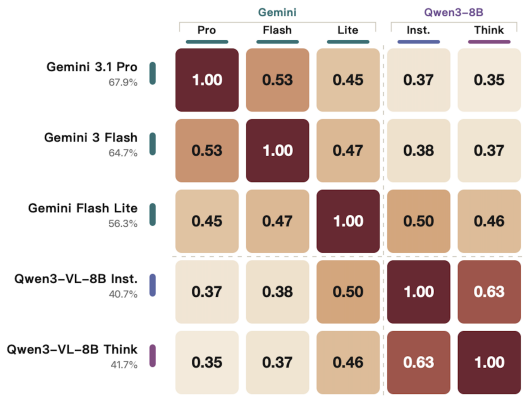


Figure 6 | **Jaccard overlap of model errors on EgoTools-Bench.** This highlights shared failure patterns within and across model families.

5. Conclusion

We introduce EgoTools, a tool-centric egocentric video suite that reframes embodied video understanding around the physical instruments through which humans act, adapt, and transform the world. By combining real-world tool-use recordings, dense hierarchical annotations, long-horizon spatial grounding, and a diagnostic QA benchmark, EgoTools exposes a key limi-

tation of current multimodal video models: recognizing activities is not sufficient to understand how tools mediate action, causality, procedure, and state change. At the same time, supervised fine-tuning on EgoTools-derived instruction data improves Qwen3-VL-8B-Instruct from 50.0% to 60.9% on EgoTools-Bench, demonstrating that the proposed data provides actionable supervision for tool-centric reasoning. Nevertheless, a substantial gap to human performance remains, particularly in grounding tool-mediated interactions in precise visual evidence. Ultimately, EgoTools pushes egocentric video understanding beyond recognizing what people do toward understanding how intelligence operates through tools, opening a path toward embodied AI systems that can observe, reason, and assist in the physical world.

References

[1] 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.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

[4] 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.

[5] Leonard Barmann and Alex Waibel. 2022. Where did i leave my keys? episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1559–1567.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, and 1 others. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. *Preprint*, arXiv:2307.15818.

[7] 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 Hall, Elissa Li, Shane Moon, Nicolas Scheffer, Kirmani Ahmed, Babak Damavandi, Rakesh Wanga, and 3 others. 2025. WearVQA: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios. *Preprint*, arXiv:2511.22154.

[8] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2024. EgoPlan-Bench: Benchmarking multimodal large language models for human-level planning. *Preprint*, arXiv:2312.06722.

[9] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. EgoThink: Evaluating first-person perspective thinking capability of vision-language models. *Preprint*, arXiv:2311.15596.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and 1 others. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55.

[11] Shangzhe Di and Weidi Xie. 2024. Grounded question-answering in long egocentric videos. *Preprint*, arXiv:2312.06505.

[12] Kuan Fang, Yuke Zhu, Animesh Garg, Andrey Kurenkov, Viraj Mehta, Li Fei-Fei, and Silvio Savarese. 2018. Learning task-oriented grasping for tool manipulation from simulated self-supervision. *Preprint*, arXiv:1806.09266.

[13] 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.

[14] Hengjian Gao, Kaiwei Zhang, Shibo Wang, Mingjie Chen, Qihang Cao, Xianfeng Wang, Yucheng Zhu, Xiongkuo Min, Wei Sun, Dandan Zhu, and Guangtao Zhai. 2026. LifeEval: A multimodal benchmark for assistive AI in egocentric daily life tasks. *Preprint*, arXiv:2603.00490.

[15] 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.

[16] Yuanyuan Gao, Hao Li, Yifei Liu, Xinhao Ji, Yuning Gong, Yuanjun Liao, Fangfu Liu, Manyuan Zhang, Yuchen Yang, Dan Xu, and 1 others. 2026. Holi-spatial: Evolving video streams into holistic 3d spatial intelligence. *arXiv preprint arXiv:2603.07660*.

[17] Google DeepMind. 2025. Gemini 3 flash model card. https://deepmind.google/models/model-cards/gemini-3-flash. Published December 2025.

[18] Google DeepMind. 2026. Gemini 3.1 flash-lite model card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/. Published March 2026.

[19] Google DeepMind. 2026. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Published February 2026.

[20] 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. *Preprint*, arXiv:2110.07058.

[21] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, and 1 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*, pages 19383–19400.

[22] 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. *Preprint*, arXiv:2403.11289.

[23] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. 2022. EgoTaskQA: Understanding human tasks in egocentric videos. *Preprint*, arXiv:2210.03929.

[24] 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.

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

[26] Chuhan Li, Ruilin Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. 2026. Learning situated awareness in the real world. *Preprint*, arXiv:2602.16682.

[27] Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. 2025. Learning precise affordances from egocentric videos for robotic manipulation. *Preprint*, arXiv:2408.10123.

[28] Hao Li, Minghan Qin, Zhengyu Zou, Diqi He, Xinhao Ji, Bohan Li, Bingquan Dai, Dingewn Zhang, and Junwei Han. 2024. Langsurf: Language-embedded surface gaussians for 3d scene understanding. *arXiv preprint arXiv:2412.17635*.

[29] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635.

[30] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2024. HOI4D: A 4d egocentric dataset for category-level human-object interaction. *Preprint*, arXiv:2203.01577.

[31] 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.

[32] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *Preprint*, arXiv:2308.09126.

[33] 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.

[34] Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. 2026. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations (ICLR)*.

[35] Ye Pan, Chi Kit Wong, Yuanhuiyi Lyu, Hanqian Li, Jiahao Huo, Jiacheng Chen, Lutao Jiang, Xu Zheng, and Xuming Hu. 2026. EgoIntent: An egocentric step-level benchmark for understanding what, why, and next. *Preprint*, arXiv:2603.12147.

[36] 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.

[37] 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. *Preprint*, arXiv:2503.13646.

[38] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. 2024. AffordanceLLM: Grounding affordance from vision language models. *Preprint*, arXiv:2401.06341.

[39] Meiying Qin, Jake Brawer, and Brian Scassellati. 2023. Robot tool use: A survey. *Frontiers in Robotics and AI*, 9.

[40] 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.

[41] Francesco Ragusa, Michele Mazzamuto, Rosario Forte, Irene D'Ambra, James Fort, Jakob Engel, Antonino Furnari, and Giovanni Maria Farinella. 2025. Ego-EXTRA: Video-language egocentric dataset for EXpert-TRAinee assistance. *Preprint*, arXiv:2512.13238.

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, and 1 others. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

[43] Daniel Seita, Yufei Wang, Sarthak J. Shetty, Edward Yao Li, Zackory Erickson, and David Held. 2022. ToolFlowNet: Robotic manipulation with tools via predicting tool flow from point clouds. *Preprint*, arXiv:2211.09006.

[44] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *Preprint*, arXiv:2203.14712.

[45] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J. Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, and 2 others. 2023. RoboVQA: Multimodal long-horizon reasoning for robotics. *Preprint*, arXiv:2311.00899.

[46] 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.

[47] 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.

[48] 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.

[49] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan 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.

[50] 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.

[51] 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.

[52] Junbin Xiao, Shenglang Zhang, Pengxiang Zhu, and Angela Yao. 2026. Ego-grounding for personalized question-answering in egocentric videos. *Preprint*, arXiv:2604.01966.

[53] 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.

[54] 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.

[55] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, 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.

[56] Lei Yao, Yong Chen, Yuejiao Su, Yi Wang, Moyun Liu, and Lap-Pui Chau. 2026. HAMMER: Harnessing MLLM via cross-modal integration for intention-driven 3d affordance grounding. *Preprint*, arXiv:2603.02329.

[57] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. 2025. SeqAfford: Sequential 3d affordance reasoning via multimodal large language model. *Preprint*, arXiv:2412.01550.

[58] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. Llava-video: Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

[59] Zixin Zhang, Kanghao Chen, Xingwang Lin, Lutao Jiang, Xu Zheng, Yuanhuiyi Lyu, Litao Guo, Yinchuan Li, and Ying-Cong Chen. 2025. PhysToolBench: Benchmarking physical tool understanding for MLLMs. *Preprint*, arXiv:2510.09507.

[60] Wenqi Zhou, Kai Cao, Hao Zheng, Yunze Liu, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio Mayol-Cuevas, Fan Zhang, Weizhe Lin, and Junxiao Shen. 2025. X-LeBench: A benchmark for extremely long egocentric video understanding. *Preprint*, arXiv:2501.06835.

[61] 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.

Appendix Contents

A. Contributors and Acknowledgments

Contributors are grouped according to their primary areas of contribution. The first four authors contributed equally through distinct core responsibilities spanning project coordination, data and benchmark development, model training, and spatial annotation.

A.1. Core Contributors

- **Shulin Tian:** Led the overall organization and coordination of the project, contributed substantially to model evaluation, and led manuscript writing and preparation.
- **Junsu Kim:** Led data collection and benchmark construction from the early stages of the project, contributed substantially to model evaluation, and led manuscript writing and preparation.
- **Shuai Liu:** Led supervised fine-tuning, LoRA-based refinement, and the associated model evaluation.
- **Hao Li:** Led the development of the 3D annotation and spatial processing pipeline, including scene reconstruction, long-horizon object tracking, 3D trajectory construction, and spatial QA generation, and contributed to supervised fine-tuning and model evaluation.

A.2. Additional Author Contributions

Data Collection and Benchmark Construction.

- **Yeongon Kim:** Contributed to egocentric video data collection, benchmark construction, and model evaluation.
- **Sihan Li:** Contributed to egocentric video data collection and benchmark construction.
- **Zhe Yang:** Contributed to egocentric video data collection and benchmark construction.

Data Collection.

- **Feiyu Li:** Contributed to egocentric video data collection.
- **Jialin Wu:** Contributed to egocentric video data collection.
- **Yichi Zhang:** Contributed to egocentric video data collection.
- **Wenhui Wang:** Contributed to egocentric video data collection.
- **Runmao Yao:** Contributed to egocentric video data collection.
- **Yuhao Dong:** Contributed to egocentric video data collection.
- **Zhaoxi Chen:** Contributed to egocentric video data collection.

Model Evaluation.

- **Yujiao Shen:** Contributed to model evaluation.

Recording Infrastructure.

- **Fangzhou Hong:** Provided the HOMIE recording platform and supporting infrastructure for multimodal egocentric data acquisition.

A.3. Advising and Supervision

- **Ziwei Liu:** Supervised the project, provided overall research guidance, and served as the corresponding author.
- **Antonino Furnari:** Advised the project and provided research feedback.
- **Jingkang Yang:** Advised the project and provided research feedback.
- **Hongyuan Zhu:** Advised the project and provided research feedback.

A.4. Additional Data Contributors

The following contributors supported egocentric video data collection and are not listed as manuscript authors:

- **Jewoo Park**
- **Jeongwon Yoo**
- **Yulgyeom Kim**
- **Yuan Fei**
- **Haiheng Liu**
- **Gordon Chen**
- **Jiarui Zheng**
- **Junhan Zhu**

We thank all participants, annotators, reviewers, and infrastructure contributors whose efforts supported the collection, annotation, quality control, and release preparation of EgoTools.

B. Ethics

EgoTools was constructed with attention to informed participation, permitted data use, privacy protection, and responsible public release. We provide an anonymized summary of the consent, data-use, and release safeguards applied during data collection, annotation, benchmark construction, model training, and dataset distribution.

Voluntary participation and informed consent. Participation in egocentric video recording, narration, and annotation was voluntary. Participants received no monetary compensation for contributing recordings, narrations, or annotations. Before contributing data, participants were informed about the purpose of the project, the types of data to be collected, the intended research uses, and the possibility that cleared materials would be publicly released for research purposes. Participants provided consent covering the applicable data collection, annotation, research-use, and release activities.

Institutional ethics review. No formal IRB or institutional ethics approval was sought for the EgoTools data collection. We explicitly report the absence of formal ethics review and describe below the informed-consent, privacy, anonymization, and release safeguards applied throughout the project.

Consent and permitted use. Participants contributed egocentric videos, narrations, and associated annotations for research on physical tool-use reasoning, egocentric video understanding, benchmark construction, grounded video-language understanding, multimodal model training, and dataset release. Within the applicable consent and data-use scope, derived materials may include corrected English narrations, dense hierarchical captions, benchmark question–answer pairs, timestamps, grounding annotations, object trajectories, and other research annotations generated from the contributed recordings.

Collected data. Depending on the participant role and recording setup, internally collected materials may include raw fisheye videos, rectified egocentric videos, synchronized audio, spoken narrations, corrected English text narrations, dense hierarchical captions, benchmark QA annotations, clip metadata, timestamp annotations, grounding-oriented focal-tool annotations, depth or pose information, and selected trajectory or object-state annotations.

Public release scope. The public release is restricted to materials covered by the applicable participant consent, release authorization, anonymization procedures, and privacy review. The release focuses on rectified egocentric videos, dense hierarchical captions, corrected English narrations, benchmark QA pairs, clip metadata, timestamps, split information, and selected finalized grounding annotations. Original narration audio and selected 3D trajectory assets are released only when covered by the corresponding release authorization and privacy review. Raw fisheye videos are not publicly released.

Privacy and release safeguards. Released metadata excludes names, contact information, institution-specific identifiers, public profiles, and exact per-video contribution counts. Sensitive, private, or accidentally captured content is reviewed before release, and materials deemed unsuitable for public distribution are excluded. Additional grounding-oriented annotations and 3D trajectory assets are released incrementally only after annotation finalization and the corresponding authorization and privacy checks.

Participant rights and data removal. Before public release, participants may request the exclusion of data associated with their participation in accordance with the project release policy. Requests received after release are handled according to the applicable consent terms, dataset license, and release procedures.

Research-use restrictions. EgoTools is distributed under a research-use license specifying permitted uses, redistribution conditions, and usage restrictions. The dataset is not intended for participant identification, biometric profiling, surveillance, or other uses that conflict with the original consent and release scope.

C. Dataset Collection and Contributor Information

C.1. Data Collection Settings

EgoTools uses two complementary collection settings to balance procedural control with natural variation in real-world tool use. We provide representative anonymized examples

below.

Structured task-guided recording. In a structured task-guided session, the data collection team prepares a task goal and a sequence of key procedural steps, while allowing the participant to perform the manipulation naturally. For example, in a kitchen recording, a participant may be asked to prepare a simple pan-cooked dish by washing ingredients, cutting them on a board, heating a pan, spreading oil, manipulating food with a spatula or chopsticks, and transferring the result to a plate. The participant is not scripted at the frame level, but the task sequence ensures that the recording contains clear procedural boundaries, repeated hand–tool–object interactions, and observable state changes such as raw ingredients being cut, food being cooked, or tools being cleaned for later use. Such sessions are useful for benchmark construction because they provide temporally coherent clips with well-defined goals and visually grounded state transitions.

Open-ended participant-driven recording. In an open-ended participant-driven session, the participant is given only a broad activity goal and completes it in their own manner. For example, a participant may be asked to perform an unscripted household, craft, laboratory, or workshop activity using the tools they consider appropriate. This setting captures natural variation that is difficult to script, such as choosing a narrower tool because the target opening is small, replacing one tool with another after an unsuccessful attempt, or changing the manipulation strategy as the target object bends, tears, loosens, or becomes unstable. These recordings complement the structured sessions by exposing models to spontaneous tool choice, substitution, recovery, and state-dependent manipulation.

C.2. Anonymized Contributor Statistics

To preserve anonymity, we report aggregate role and background statistics for contributors from whom complete information was collected. These statistics cover a documented subset of the broader author and contributor pool and are not intended to enumerate every manuscript author or every individual acknowledged in Appendix A. We do not report names, contact information, institutional identifiers, public profiles, or exact per-video contribution counts.

Contributor roles. Contributors may participate in one or more roles, including video participant, narrator, QA annotator, QA reviewer, data curator, grounding annotator, or trajectory annotator. Since several contributors played multiple roles, the role counts in Table 6 are not mutually exclusive.

Table 6 | **Summary of anonymized contributor roles.** Contributors may appear in multiple roles.

Role	# Contributors	Main Responsibility
Participant / narrator	16	Record egocentric videos and provide tool-centric narrations
QA annotator	10	Write human-crafted benchmark QA pairs
QA reviewer	2	Review answer validity, visual evidence, and distractor quality
Data curator	3	Curate videos, manage annotation progress, and organize benchmark data
Grounding annotator	1	Annotate focal-tool grounding for selected instances
Trajectory annotator	2	Construct or verify trajectory annotations for selected clips

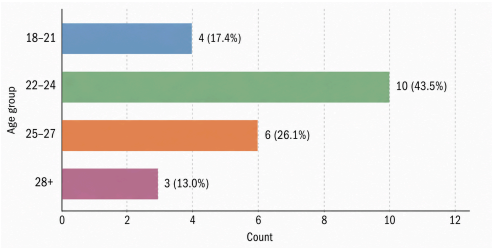
Anonymization policy. These aggregate statistics are intended to document the range of contributor backgrounds and project roles without enabling linkage to real identities or specific released recordings. We do not release names, institutions, contact information, public profiles, or exact per-video contribution counts.

Table 7 | **Summary of anonymized contributor backgrounds.** Panel (a) summarizes aggregate degree and field distributions; panel (b) shows the aggregate age-group distribution. We report only aggregate statistics and do not link age information to individual contributors.

(a) Degree and field

Degree		Field	
Category	**#**	**Category**	**#**
Undergrad.	12	Computer Science	12
Master's	5	Electrical Eng.	7
PhD	5	Biology	2
Postdoc	1	Chemistry	1
		Film / Media	1

(b) Age group



D. Preprocessing and Annotation Protocols

D.1. View Rectification Settings

Recording device and modalities. EgoTools videos are recorded with HOMIE, a head-mounted multimodal recording platform that captures four synchronized fisheye camera streams together with audio and auxiliary motion and synchronization signals. These raw modalities provide the source data for downstream view rectification, temporal alignment, and selected geometry-related processing.

Canonical view construction. For annotation, instruction-data construction, benchmark construction, model evaluation, and public release, we use the rectified front-left RGB stream as the canonical egocentric view. The rectification process uses synchronized views from the same recording to produce a spatially normalized first-person representation while preserving the hand–tool–object interactions and surrounding workspace required for tool-use understanding.

Figure 7 shows a representative rectification example. The target and reference views are processed using fixed view parameters to produce the canonical output. For the displayed example, the output resolution is 1024×1024, the zoom factor is 0.78125, the field of view is 90.0°, and the pitch correction is -5.0°, with yaw and roll set to 0.0°. The values shown in the figure correspond to the displayed example.

The resulting videos are single-view egocentric streams with a resolution of 1024×1024 at 20 FPS and synchronized audio. The same canonical video representation is used for annotation, instruction-data construction, benchmark construction, model evaluation, and the publicly released video assets.

Release scope. The public release includes the rectified canonical videos rather than the original raw fisheye camera streams. We report a representative rectification configuration and the output specifications used in our data processing, while raw fisheye streams and internal preprocessing utilities are not included in the release.

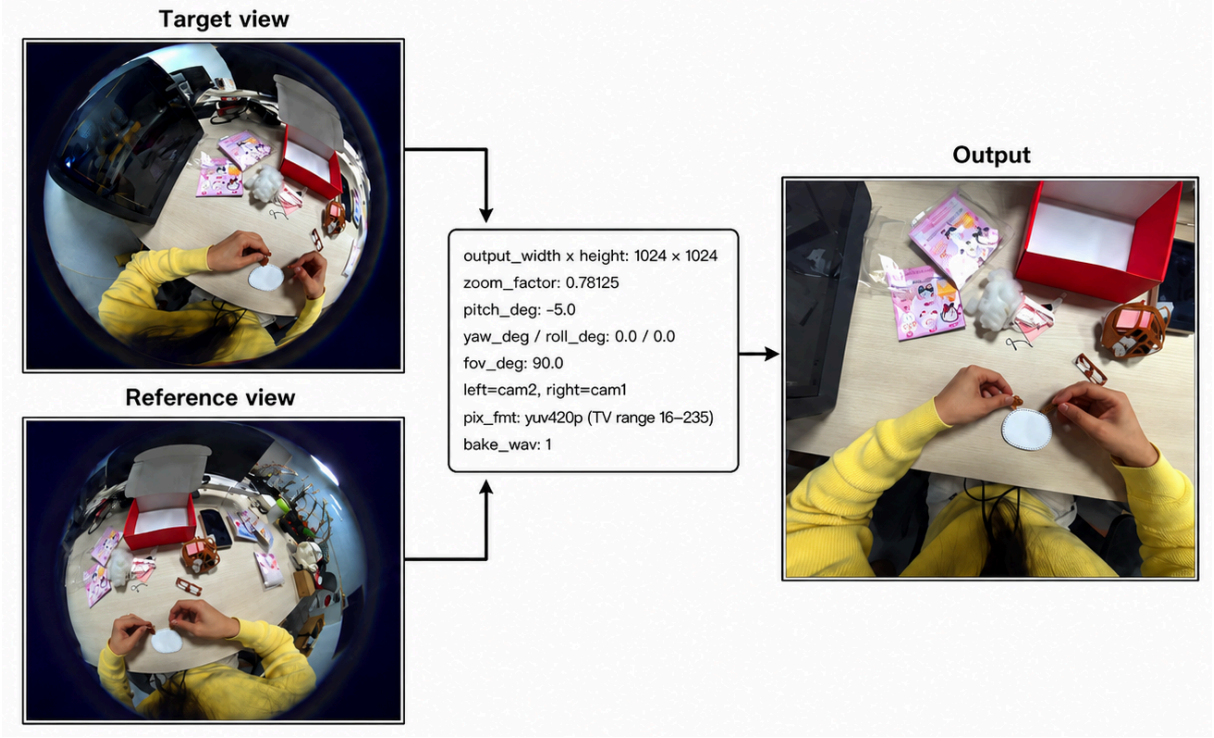


Figure 7 | **Rectification setting example.** Given synchronized target and reference views from the same recording, the rectification process applies view parameters to produce the spatially normalized canonical egocentric frame used for annotation, training, evaluation, and release. The processed view reduces fisheye distortion while preserving hand–tool–object interactions and workspace context.

D.2. Narration Annotation Guidelines

Participants provide narrations for meaningful tool-use events rather than densely describing every visible action. The goal is to capture which tool is being used, what target is being acted on, why the selected tool is appropriate in the current context, and what alternative tool may be relevant when applicable.

Annotation interface and procedure. The annotation interface allows participants or data collectors to watch rectified egocentric videos, identify relevant timestamps, and provide tool-centric narrations. Narrations are associated with temporally localized video segments and are written or corrected with reference to the visible hand–tool–object interaction.

The narration guidelines encourage annotators to specify the selected tool, a relevant alternative when applicable, the target object, the performed action, and the physical or procedural rationale for the tool choice. Narrations are free-form and need not follow a rigid sentence structure, but the following template is provided as a reference:

```
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].
```

Narration targets. Narrations focus on tool-use moments that involve meaningful decisions or state-dependent behavior, including selecting a tool, switching between tools, adapting manipulation to a changing object state, preparing a tool for a later step, recovering from an unsuccessful attempt, or using a tool in a physically appropriate manner. Participants are not

required to narrate every visible action.

Narration language and normalization. Source narrations may be provided in English, Korean, or Chinese. Korean and Chinese narrations are transcribed and translated into English, and all narrations are subsequently normalized and corrected by bilingual or English-proficient project members when necessary. The corrected English text is paired with the corresponding video segment and used as the primary narration annotation signal for instruction-data construction.

Language tags retain information about the original narration language. Original narration audio is preserved as an internal dataset asset and may be included in the public release only when covered by the applicable release authorization and privacy review.

Benchmark language and evidence separation. All final EgoTools-Bench questions and answer choices are written in English. When benchmark construction relies on auxiliary narrations or rough temporal descriptions originally provided in Korean or Chinese, these materials are translated and normalized into English before QA construction or verification. Final benchmark items are reviewed for linguistic clarity, consistency with timestamped visual evidence, single-best-answer validity, and distractor quality.

Tool-centric narrations, corrected narration text, and narration audio are not provided to models during EgoTools-Bench evaluation. They are also not supplied to benchmark annotators as privileged answer evidence; benchmark questions must remain answerable from the corresponding video and question context.

Grounding-oriented annotations. Narrations may be associated with focal tools or target objects. For selected instances, annotators provide 2D grounding points on corresponding keyframes, linking textual mentions to visible object instances. These grounding annotations provide initialization cues for long-horizon tracking and support focal-tool localization, trajectory construction, and grounded video-language training.

Grounding annotations are available only for selected instances and are progressively released as annotation finalization, release authorization, and privacy review are completed.

Examples.

- I am using the narrow chopsticks rather than the spoon to reach the jam inside the jar because the spoon is too wide to fit through the opening.
- I am using the flat spatula rather than the chopsticks to lift the egg because the broad surface can support the soft egg without tearing it.
- I am using the pipette rather than the larger container to transfer a small amount of liquid because the pipette provides finer volume control.

E. Dataset Statistics and Release

E.1. Dataset and Benchmark Statistics

EgoTools contains 646 egocentric videos totaling 100.37 hours across seven tool-use domains: kitchen, classroom, research laboratory, repair workshop, craft, office, and household. Throughout the main paper, we refer to this corpus as approximately 100 hours. The corpus additionally

Table 8 | **Summary of EgoTools dataset and benchmark statistics.** The benchmark-reserved source videos are disjoint from all videos used to construct the instruction-tuning corpus.

Statistic	Value
Total EgoTools videos	646
Total video duration	100.37 hours
Hierarchical captions	361,332
Tool-centric narrations	6,519
Benchmark-reserved source duration	40.34 hours
Benchmark QA pairs	1,000
Human-crafted QA pairs	900
Human-verified spatial QA pairs	100

Table 9 | **Domain-level composition of EgoTools-Data and EgoTools-Bench.** The corpus spans all seven collection domains, whereas the benchmark is strongly long-tailed and was not designed to be domain-balanced.

	EgoTools-Data				EgoTools-Bench
Domain	**# Videos**	**Hours**	**# Captions**	**# Narrations**	**# QAs**
Kitchen	276	32.37	116,532	2,113	837
Classroom	74	6.86	24,696	467	15
Research Lab	61	30.92	111,312	1,887	99
Repair Workshop	84	12.34	44,424	763	16
Craft	67	10.17	36,612	657	23
Office	30	2.83	10,188	213	6
Household	54	4.88	17,568	419	4
Total	**646**	**100.37**	**361,332**	**6,519**	**1,000**

contains 361,332 hierarchical captions and 6,519 tool-centric narrations. From these recordings, we reserve 40.34 hours of high-quality, tool-dense source videos exclusively for benchmark construction and evaluation.

Domain-level composition. Table 9 reports the distribution of the complete EgoTools-Data corpus and the corresponding domain allocation of EgoTools-Bench. Corpus statistics describe the full data collection, whereas benchmark QA counts describe the source domain assigned to each of the 1,000 benchmark questions.

The domain distribution highlights the distinction between corpus coverage and benchmark composition. EgoTools-Data contains substantial recordings from all seven domains, including approximately 31 hours of research-laboratory activity and more than 12 hours of repair-workshop activity. EgoTools-Bench, however, is dominated by Kitchen questions and should not be interpreted as a domain-balanced evaluation set. The benchmark was constructed primarily to cover complementary tool-use reasoning capabilities, annotation quality, and challenging tool-mediated interactions rather than to support statistically balanced comparisons across collection domains. Domain-level results for the smallest benchmark slices would therefore be unstable and are not used for primary performance claims.

Reasoning-track distribution. EgoTools-Bench contains **1,000** 8-way multiple-choice question–answer pairs, including **900** human-crafted questions and **100** human-verified spatial questions generated from the 3D annotation pipeline. Across the four research-facing tracks, the benchmark includes 363 Affordance & Causality questions, 236 Perception & Grounding questions, 222 Procedural Dynamics questions, and 179 Spatial Reasoning questions. This distribution reflects the benchmark's emphasis on tool selection, application, and physical effects while maintaining coverage of perceptual grounding, procedural understanding, and spatial hand–tool–object relations.

Evaluation-only benchmark partition. All benchmark source videos, clips, questions, answers, and derived annotations are reserved exclusively for evaluation. No clip, caption, narration, synthetic QA, or other annotation derived from a benchmark-reserved source video is included in the instruction-tuning corpus.

E.2. Release and Privacy Considerations

The EgoTools release is scoped to research use and includes only materials covered by the applicable participant consent, release authorization, anonymization procedures, and privacy review. The release provides processed rectified egocentric videos rather than the original raw fisheye recordings, together with the annotations and metadata required for benchmark evaluation and dataset documentation. Additional consent and privacy safeguards are described in Appendix B.

Initial release components. The initial release package includes the following cleared components:

- Rectified egocentric video clips;
- Dense hierarchical captions;
- Corrected English tool-centric narrations;
- EgoTools-Bench questions and evaluation annotations;
- Clip metadata and timestamps;
- Source-video-disjoint training and benchmark split information; and
- Dataset documentation and benchmark evaluation metadata.

Incrementally released components. The following components are released incrementally as their annotation, authorization, and privacy-review procedures are completed:

- Original narration audio for consent-cleared examples;
- Selected finalized focal-tool grounding annotations;
- Selected object-state and trajectory annotations; and
- 3D trajectory visualizations for selected clips.

The availability of these components may differ across examples because grounding, audio, and trajectory assets require additional annotation-finalization and release-review procedures.

Excluded components. Raw fisheye camera streams are not publicly released. The release also excludes personally identifying metadata, uncleared recordings or audio, sensitive or accidentally captured content, internal identity mappings, exact per-video contributor counts, and internal preprocessing utilities.

Privacy safeguards. Names, contact information, institution-specific identifiers, public-profile information, and other personally identifying information are removed or withheld from released metadata. Sensitive, private, or accidental content is reviewed before release, and materials deemed unsuitable for public distribution are excluded. Contributors may request exclusion of data associated with their participation in accordance with the release policy and applicable consent terms.

Access conditions and license. EgoTools is distributed under a research-use license specifying permitted uses, redistribution conditions, and usage restrictions. Privacy-sensitive or conditionally released components are handled according to their applicable authorization and privacy-review outcomes. The dataset is not intended for participant identification, biometric profiling, surveillance, or other uses inconsistent with the original collection and release scope.

Limitations. Although EgoTools-Data spans seven real-world tool-use domains, it does not exhaustively cover all tools, environments, participant populations, cultural practices, or professional procedures. The benchmark distribution is strongly concentrated in Kitchen videos and should not be treated as representative of the domain distribution of the full corpus. Grounding, narration-audio, and 3D trajectory components are available only for examples that have passed the corresponding annotation-finalization, release-authorization, and privacy-review procedures.

F. Quality Control and Normalization

F.1. Fix-First Quality Control

Raw human annotations are processed under a fix-first quality-control policy that prioritizes repairing valid annotations before discarding them. Items are retained whenever deterministic or minimal model-assisted edits can preserve the original annotator intent without introducing ambiguity.

Structural audit. Each item is evaluated using a 22-flag binary audit covering defects such as placeholders, missing fields, malformed options, duplicate entries, first-person leakage, personally identifiable information, answer–option mismatch, inappropriate full-video clip usage, and distractor anomalies.

Filtering policy. Content-empty submissions and verbatim duplicates are removed. Repairable defects are corrected through deterministic rules or minimal model-assisted edits. Cases for which the intended meaning or unique answer cannot be established with high confidence are routed to human review rather than being automatically accepted.

Model-assisted repair. We use Gemini-2.5-Flash-Lite with temperature 0 for model-assisted repair. The repair prompt instructs the model to perform the smallest meaning-preserving edit possible. Typical edits include replacing personally identifying names with role-based references, rewriting first-person expressions into third-person form, correcting surface-level grammar, and

completing truncated answer text only when the intended content is unambiguous. Edits that could alter the semantic content or answer validity are routed to human review.

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

Eight-choice normalization. All benchmark items are normalized to exactly eight answer options. We dispatch each item according to its current distractor count: items with fewer than seven distractors are augmented, items with exactly seven distractors are validated, and items with more than seven distractors are reduced to seven plausible alternatives. This process produces a consistent 8-way multiple-choice format while preserving the original correct answer and question intent whenever possible.

Anti-shortcut checks. Each normalized item and the resulting benchmark corpus are subjected to deterministic anti-shortcut checks before finalization. Item-level checks detect meta-options such as "all of the above" or "none of the above," duplicate or near-duplicate options, token-set permutations of the correct answer, substring or superstring containment, and excessive question–answer lexical overlap. Corpus-level checks additionally examine answer-position imbalance and systematic answer-length bias. Failed items are routed to bounded regeneration or human review, with detected violations provided as feedback for revising the answer options.

Length-bias check. For each item, we compare the character length of the correct answer with the distribution of distractor lengths. 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 satisfying $|z| > 1.5$ are flagged for repair. Previously cached eight-option snapshots are subjected to the same checks rather than being accepted without revalidation, ensuring that the length constraint is consistently enforced across all benchmark items. An earlier cached snapshot exhibited a corpus-level mean standardized deviation of $\bar{z} \approx +2.21$. After applying the finalized check and repair procedure, the retained benchmark has $\bar{z} \approx -0.15$, with every item satisfying $|z| \leq 1.5$.

Option shuffling. The final answer-option order is shuffled using a deterministic seed derived from each QA identifier. This reduces positional priors while keeping evaluation reproducible. Because every released benchmark item contains eight answer choices, the main benchmark tables report standard multiple-choice accuracy with a consistent chance baseline of 12.5%.

G. Additional Related Work

Broader egocentric video reasoning. Egocentric video benchmarks have expanded from action and object recognition toward episodic memory, planning, temporal localization, situated reasoning, and assistance. QAEgo4D and GroundVQA study episodic-memory QA and temporally grounded QA in long egocentric videos [5, 11]. EgoTaskQA evaluates task-oriented questions about dependencies, effects, goals, and beliefs [23]; EgoSchema and EgoTempo emphasize long-context and temporal reasoning [32, 37]; and EgoPlan-Bench evaluates planning from first-person videos [8]. More recent benchmarks examine first-person thinking, life-log and real-time assistance, expert–trainee support, situated awareness, multi-agent egocentric QA, intent understanding, personalized grounding, and AR or life-logging environ-

ments [9, 14, 24, 26, 35, 41, 47, 52, 55, 60]. These resources substantially broaden egocentric evaluation around memory, temporal context, planning, and user assistance. However, they generally do not isolate the physical reasoning required to determine why a particular tool is appropriate, which substitute is feasible, or how manipulation should adapt as the target object changes state.

Instructional, embodied, and robotic interaction benchmarks. Instructional-video datasets provide useful comparisons because they frequently contain tools, ingredients, techniques, and alternative procedures. VidDetours retrieves procedural detours from how-to videos, while StepDiff identifies differences between instructional clips [1, 33]. Their primary focus, however, is procedural retrieval or comparison rather than physical tool-use reasoning grounded in a continuous first-person task. Embodied interaction benchmarks provide another related but distinct setting. OpenEQA evaluates open-vocabulary embodied question answering in real-world environments [31], RoboVQA provides large-scale video-language supervision from robotic demonstrations [45], and RoboCasa365 studies simulated kitchen manipulation across diverse tasks and environments [34]. These resources offer important scope contrasts, but they do not provide the same combination of natural human egocentric video, tool-centric narration, narration-linked focal-tool grounding, and explicit QA over tool choice, substitution, ordered 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 activities such as sweeping and hammering [12], while ToolFlowNet predicts dense tool motion for manipulation tasks such as scooping and pouring [43]. Broader surveys and manipulation methods study robot tool use, affordance-based manipulation, precise affordance grounding, and physically grounded policies [6, 15, 27, 39, 54]. Recent multimodal benchmarks and methods further evaluate whether MLLMs and VLMs can recognize affordances, infer physical constraints, and ground tool-related interactions [22, 38, 57, 59, 56]. These studies reveal persistent weaknesses in reasoning about tool functions, physical constraints, and multi-step interactions. Nevertheless, many evaluations are based on static images, synthetic or reconstructed 3D scenes, short interactions, or constrained robotic settings. EgoTools-Bench is complementary: it evaluates physical tool-use reasoning from natural first-person videos in which models must interpret real human tool selection, substitution, manipulation, and adaptation over time.

Wearable assistants. Wearable and egocentric assistant benchmarks are especially relevant because they require models to answer situated questions from the user's viewpoint. WearVQA studies visual question answering for wearable devices [7], while visual-intention grounding evaluates whether models can infer user needs and intentions from egocentric observations [46]. These settings motivate first-person assistance but do not specifically center tool-mediated physical reasoning. EgoTools-Bench therefore complements wearable-assistant research by focusing on questions that require understanding the functional role of tools, feasible substitutes, temporally ordered manipulation, and adaptation to changing object states.

H. Training and Evaluation Details

H.1. Single-Stage Training Procedure

We initialize EgoTools-8B from Qwen3-VL-8B-Instruct and train it in a single stage of full supervised fine-tuning on the EgoTools-derived instruction corpus. All training videos remain

disjoint from EgoTools-Bench at the source-video level.

Training mixture. The corpus contains 184,679 instruction examples derived exclusively from the EgoTools training pool, each labeled with the capability it supervises. Table 10 reports the breakdown. Tool-centric supervision covers the four EgoTools-Bench abilities directly: why a tool suits a goal or material (AC), which tool is in use and what local state it is in (PG), how a task advances within a clip (PD), and how hands, tools, and objects are arranged in space (SR). Episode-level supervision asks about action ordering, overall activity, and the goal of a whole recording. General video QA contributes mixed episode-level questions that are not targeted at a single ability. Descriptive supervision through dense captioning and narration completion teaches the model to verbalize first-person visual evidence outside a question format, and single-image open-ended QA covers free-form answering from one frame. By answer format, the mixture contains 170,595 multiple-choice examples, 5,000 short open-ended examples, and 9,084 dense-captioning and narration-completion examples; answer letters in the next-action subset are balanced across the four positions at 1,250 examples each.

The next-action and single-image open-ended examples are generated from EgoTools training videos and their caption timelines; no video, image, or QA annotation from EgoSchema, EgoPlan-Bench, or EgoThink is used. Because these two formats mirror how EgoPlan-Bench and EgoThink pose their questions, a subset of their question strings coincides with generic benchmark question templates; media overlap with the benchmark reference set is zero, and we treat media disjointness as the operative leakage criterion. We accordingly attribute the EgoPlan and EgoThink results in Table 11 to answer-format coverage rather than to general transfer.

Optimization. We perform full-parameter supervised fine-tuning of the language-model component of Qwen3-VL-8B-Instruct, while keeping the vision encoder and multimodal projector frozen. Thus, "Full SFT" refers to full-parameter updating of the language-model component rather than updating every module in the multimodal architecture.

We optimize using AdamW with a learning rate of 2.3×10^{-6}, a constant learning-rate schedule, and no warmup. Training is performed for one epoch with a per-device batch size of 1, gradient accumulation over 16 steps, and 8 GPUs, yielding an effective batch size of 128 and 1,443 optimizer steps. We use bfloat16 precision, DeepSpeed ZeRO Stage 3, FlashAttention, and gradient checkpointing.

Each training video is represented by 64 uniformly sampled frames at a target rate of 2 FPS, with the per-video frame count fixed to 64. We cap each image at 1,024 visual tokens and use a maximum multimodal sequence length of 8,192 tokens.

Source-video separation. The EgoTools training pool and EgoTools-Bench are partitioned before instruction-data or benchmark construction. No video, clip, caption, narration, synthetic QA pair, grounding annotation, or other derivative of a benchmark-reserved source video is used during training.

H.2. Training Data Composition

The training procedure uses 184,679 EgoTools-derived examples. Table 10 summarizes their composition by the capability each example supervises.

Supervision aimed directly at the four EgoTools-Bench abilities accounts for 37.8% of the mixture, and the four abilities are covered at comparable scale, ranging from 14,175 examples for Affordance & Causality to 22,943 for Procedural Dynamics. Spatial Reasoning receives 17,887

Table 10 | **Composition of the EgoTools-derived training data by supervised capability.** Counts describe the 184,679-example single-stage mixture, grouped by what each example supervises. The first four rows target the four EgoTools-Bench reasoning abilities directly; the remaining rows supervise episode-level understanding, general video QA, description, and open-ended answering. 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 11 | **Effect of the single-stage EgoTools training procedure.** All EgoTools-Bench results use the full quality-controlled 1,000-question benchmark. Training and benchmark videos are disjoint at the source-video level.

Model	**AC**	**PG**	**PD**	**SR**	**Overall**	**EgoSchema**	**EgoPlan**	**EgoThink**
Qwen3-VL-8B-Instruct	46.1	45.9	57.1	**55.4**	50.0	**69.0**	**42.3**	61.0
EgoTools-8B	**55.4**	**59.7**	**68.1**	44.4	**60.9**	67.6	35.2	**64.0**

examples, accounting for 9.7% of the overall training mixture. Despite this supervision, SR performance decreases after fine-tuning, suggesting that spatial reasoning remains a limitation of the current training recipe.

H.3. Checkpoint-Wise Evaluation

Table 11 summarizes the effect of EgoTools training under the same standard 64-frame MP4-based EgoTools-Bench evaluation protocol. Existing egocentric benchmarks are evaluated using their corresponding benchmark protocols: EgoSchema on the full 5,031-question set, EgoPlan-Bench on its 3,343-question validation set, and EgoThink as the macro-averaged score over its twelve single-image subtasks, judged by `gpt-4o-2024-11-20`. No videos or annotations from EgoSchema, EgoPlan-Bench, or EgoThink are used during EgoTools training.

Full supervised fine-tuning improves EgoTools-Bench accuracy from 50.0% to 60.9%, with gains on Affordance & Causality, Perception & Grounding, and Procedural Dynamics and a decrease on Spatial Reasoning. On existing benchmarks, the single-stage model reaches 67.6% on EgoSchema, within 1.4 points of the 69.0% backbone and without any separate refinement stage; an earlier 220,334-example mixture had fallen to 51.7% and required format-aligned LoRA refinement to recover.

Transfer remains benchmark-dependent. The model reaches 35.2% on EgoPlan-Bench compared with 42.3% for the original backbone, while EgoThink improves from 61.0% to 64.0%. Since the training mixture contains 5,000 next-action multiple-choice and 5,000 single-image open-ended examples, matching the answer formats those two benchmarks use, we read these

results as evidence of answer-format coverage rather than of general transfer. We therefore interpret the training results as evidence that EgoTools-derived supervision improves tool-centric egocentric reasoning, rather than as evidence of uniform transfer to every egocentric benchmark.