EgoTools: Tool-Centric Reasoning

in Real-World Egocentric Videos

Shulin Tian[1,*] **Junsu Kim**[1,2,*] **Shuai Liu**[1,3] **Hao Li**[1] **Yujiao Shen**[1] **Sihan Li**[1] **Zhe Yang**[1]
Yeongon Kim[4] **Runmao Yao**[1] **Yuhao Dong**[1] **Fangzhou Hong**[1,★] **Antonino Furnari**[5]
Jingkang Yang[3] **Hongyuan Zhu**[6] **Ziwei Liu**[1,†,★]

[1]S-Lab, Nanyang Technological University [2]KAIST [3]Synvo AI [4]Chung-Ang University
[5]University of Catania [6]A*STAR

Project Page: example.com **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 1K QA benchmark spanning diverse real-world environments and reasoning tasks.

Abstract

Ropedia

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. Yet despite strong performance on perception-oriented video tasks such as captioning

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

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 and 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 in EgoTools-Bench, validating the difficulty of the benchmark and the value of the corpus. Together, these resources establish EgoTools as a foundation for real-world egocentric tool-use understanding, paving the way for future research in this direction.

1. Introduction

Human activities in the physical world, from daily chores to professional procedures, are not only a matter of bare hands, but also depend on the skilled use of tools. For an embodied agent, understanding such activities demands more than simply recognizing visible actions and objects. It also requires the model to reason about tool–target contact, spatial coordination, procedural progress, and the physical outcome for each action. 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. In 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 on 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, where tools are often present in the video but are not foregrounded as the central unit of explanation. For previous benchmarks [8, 9, 11, 23, 32, 37], where actions may be labeled or queried at the level of “cook eggs", the spatula and pan that mediate the activity can go unmentioned, and the tools through which actions unfold are effectively invisible in the annotation. This dual absence has left a critical gap in both training signal and evaluation framework.

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-

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+	5,063	🎥	✓	✗	✗	✓	✗	✗
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	🎥⬛⊕	✓	✓	✓	✓	✓	✓

level captions to narrations that explicitly foreground affordances, procedural structure, and causal effects, as well as supplementary 3D information. This rich information jointly transforms passive video into a structured signal of 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 set up preconditions and 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 the temporal organization of 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. Together, these resources establish EgoTools as an integrated foundation for advancing real-world egocentric tool-use understanding.

We evaluate representative multimodal video models on EgoTools-Bench. Current open-source models remain clustered around the low-40% range, indicating that EgoTools-Bench provides a challenging diagnostic test for egocentric tool-use reasoning. These results show that EgoTools exposes an important gap in existing video-language models and provides a foundation for studying tool selection, procedural progress, and spatial hand–tool–object interactions.

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 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 and tool-centric narration 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 how a curated 40.34-hour subset of EgoTools-Data is used to construct and quality-control EgoTools-Bench, a 1,000-question 8-way multiple-choice diagnostic benchmark covering affordance and causality, perception and 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.

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

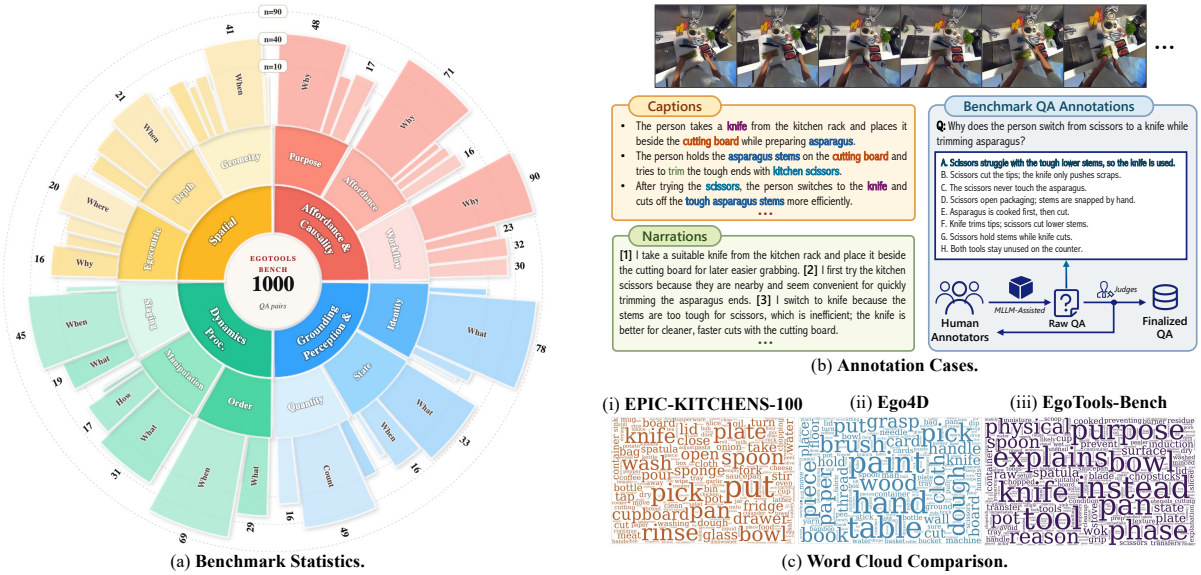


Figure 2 | **Benchmark distribution and annotation examples.** (a) Distribution of 1,000 QA pairs across four tool-use reasoning tracks and their fine-grained subtracks. (b) Example annotations, including captions, tool-centric narrations, and finalized QA pairs produced through human annotation with MLLM-assisted judging. (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.

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 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.2.

The collected and curated EgoTools 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

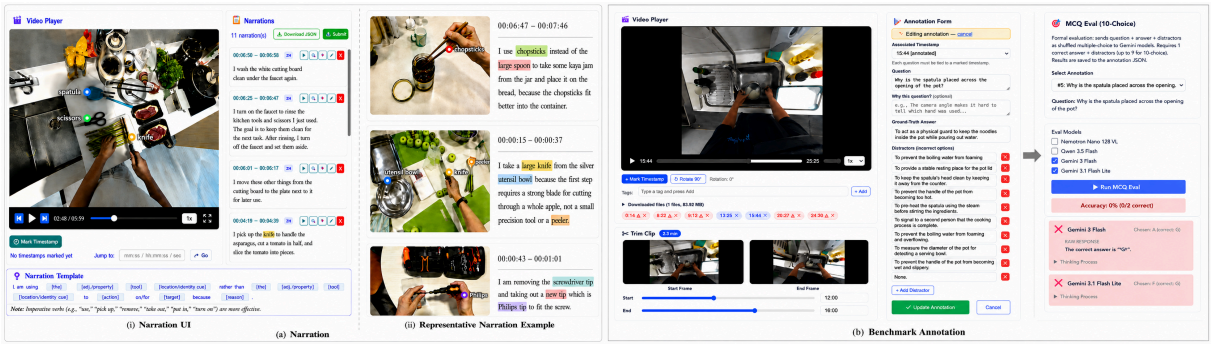


Figure 3 | **Textual annotation pipeline.** (a) Tool-centric narrations are annotated from grounded egocentric video segments. (b) Benchmark QA pairs are constructed and refined through an annotation interface with MLLM-assisted checking.

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.

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 with a reference narration template, emphasizing tool selection, usage intent, and effects on target objects or task states. For each narration, collectors annotate 2D grounding points for the mentioned tools or objects that are important to be tracked on corresponding keyframes, linking textual mentions to visual tool instances and providing initialization cues for long-horizon 3D tracking. The textual annotation interfaces are shown in Figure 3, where more details related to the narration template, 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. 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, which serves as the foundation for EgoTools-Bench construction.

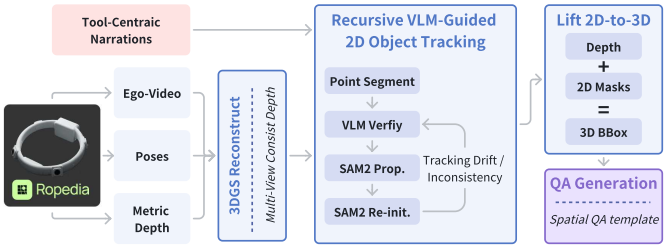


Figure 4 | **3D annotation pipeline.** Raw geometry from the HOMIE device is reconstructed with 3DGS to obtain consistent depth, then combined with tool-centric narrations for recursive VLM-guided object tracking. The resulting masks are lifted to 3D for object boxes and spatial QA generation.

Benchmark QA Construction. Based on EgoTools-Data, we select a high-quality **40.34-hour** video subset for EgoTools-Bench construction, prioritizing annotation quality, tool-use density, task diversity, and environmental coverage. 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. Items easily answered without video understanding or by

many models are hardened by replacing distractors with visually plausible, length-balanced alternatives while preserving the original question and ground-truth answer when possible. All finalized items are re-checked for a unique, visually supported, and temporally grounded correct answer. Details are in Appendix F.

3.3. 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 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. EgoTools-Bench is evaluation-only; to avoid leakage and over-counting, repeated clips from the same source video are controlled in the evaluation set. Figure 2 also compares the vocabulary distribution of EgoTools-Bench with EPIC-KITCHENS-100 and Ego4D. The word clouds show that EgoTools-Bench is more tool-dense and tool-centric, emphasizing tools, manipulated objects, actions, and state changes rather than 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 3.3; 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.

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 33.2% to 41.1%, far below the human expert score of 83.2%. The strongest open-source instruct result is Qwen2.5-Omni-7B-Instruct [53] at 41.1%, followed closely by InternVL3.5-8B-Instruct [51] at 40.8% and Qwen3-VL-8B-Instruct [2] at 40.7%. These

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	**45.2**	61.3	40.9	36.6	39.1	**46.7**	40.7
Qwen3-VL-4B-Instruct [2]	64	✗	**69.6**	42.0	**61.9**	42.2	35.7	33.9	38.0	39.2
Qwen2.5-VL-7B-Instruct [3]	64	✗	61.0	32.0	57.7	38.1	31.2	40.9	**46.7**	38.6
Qwen2.5-Omni-7B-Instruct [53]	64	✓	65.2	–	58.9	41.7	33.0	**42.6**	**46.7**	**41.1**
Qwen2-VL-7B-Instruct [50]	64	✗	63.6	39.9	60.1	41.1	30.4	38.3	37.0	38.3
InternVL3.5-8B-Instruct [51]	64	✗	63.8	–	56.0	**42.8**	34.8	36.5	45.6	40.8
InternVL3-8B-Instruct [61]	64	✗	**69.6**	–	59.4	37.3	34.8	33.9	45.6	37.5
LLaVA-OneVision-7B [25]	64	✗	63.4	–	54.9	37.9	**37.4**	31.3	37.0	36.6
LLaVA-Video-7B-Qwen2 [58]	64	✗	52.4	–	58.0	34.6	28.6	31.3	35.9	33.2
MiMo-VL-7B-SFT [48]	64	✗	54.4	35.9	60.7	37.9	26.8	39.1	39.1	36.4
Open-Source Models (Thinking)										
Qwen3-VL-8B-Thinking [2]	512	✗	**70.3**	–	62.0	**44.7**	31.3	36.5	**48.9**	41.7
Qwen3-VL-4B-Thinking [2]	512	✗	59.8	–	62.3	36.2	24.1	31.3	35.9	33.4
MiMo-VL-7B-RL [48]	64	✗	57.0	**36.6**	59.4	44.1	**44.4**	37.5	38.0	**42.3**
GLM4.1V-Thinking [49]	64	✗	66.1	25.3	**62.7**	35.9	30.9	**39.0**	43.0	36.5

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.

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 61.3% on EgoThink [9], but only 40.7% on EgoTools-Bench. Qwen3-VL-4B-Instruct shows the same gap, with 69.6% on EgoSchema, 61.9% on EgoThink, and 39.2% 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 each yield only marginal gains in overall accuracy. The gap is therefore distributed rather than localized, and we decompose it by track and model class in §4.3, then by tool-use cues and error patterns in §4.4.

4.3. Analysis

While the main results provide an overview of model performance, they do not identify which specific abilities account for the gap. We therefore analyze the results at a finer level, using the four benchmark tracks and model families to ask when models are actually grounded in the observed tool-mediated interaction and when they can rely on language priors, task context, or physical commonsense.

Perception & Grounding is the diagnostic anchor. Perception & Grounding (PG) is the least shortcut-friendly track because it asks for concrete visual facts: 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. This weakness is broad across model families. Open-source instruct models often score in the low-to-mid 30% range on PG, including Qwen2.5-Omni-7B-Instruct at 33.0%, InternVL3.5-8B-Instruct at 34.8%, Qwen3-VL-8B-Instruct at 36.6%, and LLaVA-Video-7B-Qwen2 at 28.6%. Stronger systems improve but do not remove the issue: Gemini-3.1-Pro reaches 51.7%, still 18–23 points below its Affordance & Causality, Procedural Dynamics, and Spatial Reasoning scores. This makes PG a direct test of whether a model is anchored to the exact tool-mediated interaction rather than only recognizing the surrounding activity.

Higher-level tracks distinguish how models use context. In contrast to Perception & Grounding (PG), the Affordance & Causality (AC), Procedural Dynamics (PD), and Spatial Reasoning (SR) tracks admit different but parallel sources of support: tool-function priors for AC, task-order context for PD, and physical commonsense for SR. The pattern is not that these tracks are easy, but that they expose which forms of context a model can exploit. Qwen3-VL-8B-Instruct improves over its 4B counterpart mainly on PD and SR, suggesting that additional capacity helps with temporal context and workspace geometry more than with exact grounding. Qwen2.5-Omni-7B-Instruct is only slightly stronger than Qwen2.5-VL-7B-Instruct on AC, PG, and PD, and ties it on SR, indicating that synchronized non-narration audio adds limited information for these tool-centric questions. InternVL3.5-8B-Instruct improves over InternVL3-8B-Instruct on AC and PD, while remaining unchanged on PG and SR. These track-wise differences show why EgoTools-Bench is diagnostic: it separates generic gains in video or language modeling from the specific abilities needed for counterfactual tool choice, state tracking, and hand–tool–object geometry.

Model classes fail in distinguishable ways. The model-class comparison further suggests that size, modality, and deliberation address different parts of the problem rather than one common bottleneck. The dominant contrast is proprietary versus open-source: Gemini-3 models occupy the 55–67% overall range and show the largest gains, but their weakest track remains PG, so stronger native video capacity still leaves a precise-grounding gap. Thinking-style inference changes where models fail rather than uniformly improving them. Qwen3-VL-8B-Thinking gains on AC and SR but drops on PG and PD, while GLM4.1V-Thinking performs comparably to instruct models on EgoSchema and EgoThink yet reaches only 36.5% on EgoTools-Bench. MiMo-VL-7B-RL, another thinking-style variant, provides a complementary pattern: it is the strongest open-source model overall and has the highest open-source PG score, but remains below 40% on PD and SR. These patterns suggest that the gap is best explained as an evidence-anchoring failure rather than a reasoning, scale, or modality failure alone.

4.4. Ablations and Error Analysis

Implicit tool-use questions are consistently harder. Table 3 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 4.83–7.74 points. Qwen3-VL-8B-Instruct shows the same direction with a smaller 3.97-point gap. This suggests that current models benefit substantially 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 -0.62 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 3 | **Ablation on implicit and explicit tool-use questions.** Δ denotes Explicit–Implicit.

Model	Overall	Implicit	Explicit	Δ
Gemini-3.1-Pro	**67.93**	**65.15**	**72.89**	+7.74
Gemini-3-Flash	64.72	63.13	67.96	+4.83
Gemini-3.1-Flash-Lite	56.27	54.29	59.51	+5.22
Qwen3-VL-8B-Instruct	48.52	47.18	51.15	+3.97
Qwen3-VL-8B-Thinking	41.69	42.17	41.55	-0.62

Error overlap reveals systematic failure patterns. The error map in Figure 5 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 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 the view that EgoTools-Bench errors are systematic rather than random, and that remaining failures concentrate around tool-centric video evidence such as occlusion, clutter, visually similar instruments, and brief state transitions.

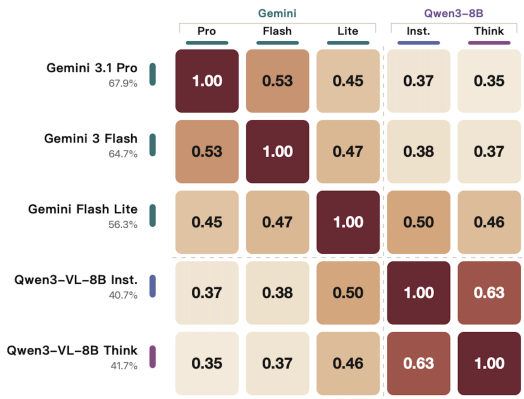


Figure 5 | **Jaccard overlap of model errors on EgoTools-Bench.** This highlights shared failure patterns within and across model families.

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 relying on language, task-order, or commonsense shortcuts when the video evidence is required.

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 trans-

form 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 limitation of current multimodal video models: recognizing activities is not enough to understand how tools mediate action, causality, procedure, and state change. Our results show that today's models still struggle to ground tool use in visual evidence, highlighting the need for future supervision and modeling approaches tailored to tool-mediated reasoning. Ultimately, EgoTools pushes egocentric video understanding beyond recognizing what people do toward understanding how intelligence operates through tools, opening a path to embodied AI that can observe, reason, and assist within 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. Authorship Statement (Work in Progress)

Project coordination and writing. Shulin Tian and Junsu Kim served as core contributors to the EgoTools project, coordinating dataset construction, benchmark design, annotation protocols, quality control, experiments, and manuscript writing.

Data collection and annotation. The broader author team contributed to egocentric video collection, tool-centric narration, QA annotation, QA review, data curation, grounding annotation, trajectory annotation, and release preparation. Contributors involved in participant recording, narration, benchmark writing, review, curation, grounding, and trajectory annotation are summarized in Appendix C.2.

System, infrastructure, and advising. The Ropedia team supported the HOMIE recording platform and spatial annotation pipeline. The S-Lab team provided infrastructure, feedback, and project-level support. Ziwei Liu supervised the project and provided overall research guidance.

B. Ethics

We provide an anonymized summary of the consent, data-use, and release safeguards used for EgoTools data collection. Participants were informed that their egocentric videos and narrations would be used for research on physical tool-use reasoning, egocentric video understanding, benchmark construction, and model training. Public distribution is restricted to materials covered by the applicable release authorization and privacy review; materials outside this scope are excluded from the release package.

Consent and permitted use. Participants contributed egocentric videos and narrations for the construction of the EgoTools dataset, including benchmark question-answer pairs and derived research annotation. Within the applicable data-use scope, the collected and derived data may be used for academic research on physical tool-use reasoning, egocentric video understanding, grounded video-language understanding, omni-modal modeling, multimodal model training, and dataset release.

Collected and released data. Depending on the participant role and recording setup, the internally collected data may include egocentric video recordings, synchronized audio, spoken narrations, corrected English text narrations, benchmark QA annotations, clip metadata, timestamp annotations, grounding-oriented focal-tool annotations, and selected trajectory or object-state annotations.

The public release is scoped to research use and includes only materials cleared for distribution. The release package focuses on rectified egocentric videos, corrected English narrations, benchmark QA pairs, metadata, and selected finalized grounding annotations. Original narration audio and selected 3D trajectory visualizations are included only when covered by the release authorization and privacy review. Raw fisheye videos will not be publicly released.

Privacy, participant rights and release safeguards. Released metadata excludes names, contact information, institution-specific identifiers, public profiles, and exact per-video contribution counts. Contributor profiles are reported only in anonymized and coarse-grained form. Before public release, contributors may request removal or exclusion of data associated with their participation according to the release policy. The final release follows applicable institutional guidance, anonymization requirements, and privacy review procedures.

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. Anonymous Contributor Profiles

To preserve anonymity, we report contributor information using randomly assigned anonymous IDs and coarse-grained attributes. We do not include names, contact information, institution-specific identifiers, public profiles, or exact per-video contribution counts. The mapping between anonymous IDs and real identities is kept only internally.

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 4 are not mutually exclusive.

Table 4 | **Summary of anonymous contributor roles.** Contributors may appear in multiple roles.

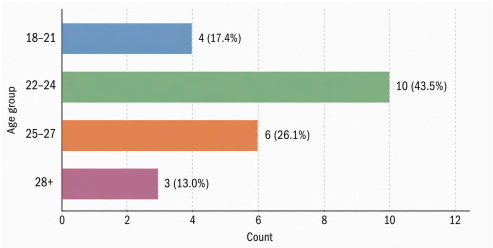
Role	# Contributors	Main Responsibility
Participant / narrator	16	Record egocentric videos and provide tool-centric narrations
QA annotator	10	Write human-crafted benchmark QA pairs
QA reviewer	2	Review answer validity, visual evidence, and distractor quality
Data curator	3	Curate videos, manage annotation 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

Table 5 | **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 contributor IDs.

(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



Anonymization policy. We report coarse contributor profiles with randomly assigned IDs and avoid names, institutions, contact information, public profiles, and exact per-video contribution counts. These profiles are intended to document the range of contributor backgrounds and roles without linking released data to real identities.

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 device that captures four synchronized fisheye camera streams together with audio and auxiliary motion/synchronization signals. These raw modalities provide the basis for downstream view rectification, temporal alignment, and selected spatial processing. For annotation, benchmark construction, and model training, we use the rectified front-left RGB stream as the canonical egocentric view while preserving the synchronized audio and original raw recordings as dataset assets.

View rectification. To obtain a more natural egocentric perspective, we apply a rectification process that adjusts zoom and pitch using a target view and a reference view to guide the viewpoint transformation. Figure 6 shows one representative rectification example, including the input views, rectification parameters, and resulting output frame. The resulting videos are single-view egocentric streams with a resolution of 1024×1024 at 20 FPS and synchronized audio.

The rectified output provides a consistent visual representation of the manipulation scene while retaining the relevant tool-use context. Compared with the raw fisheye inputs, the output frame reduces peripheral distortion and makes the spatial relations among hands, tools, target objects, and the workspace easier to inspect. In this example, the rectification uses a 1024×1024 output resolution, a zoom factor of 0.78125, a 90.0° field of view, and a -5.0° pitch correction, with yaw and roll kept at 0.0°. These rectified frames are used for visualization and annotation support; the original synchronized video data remain the source recordings for benchmark construction.

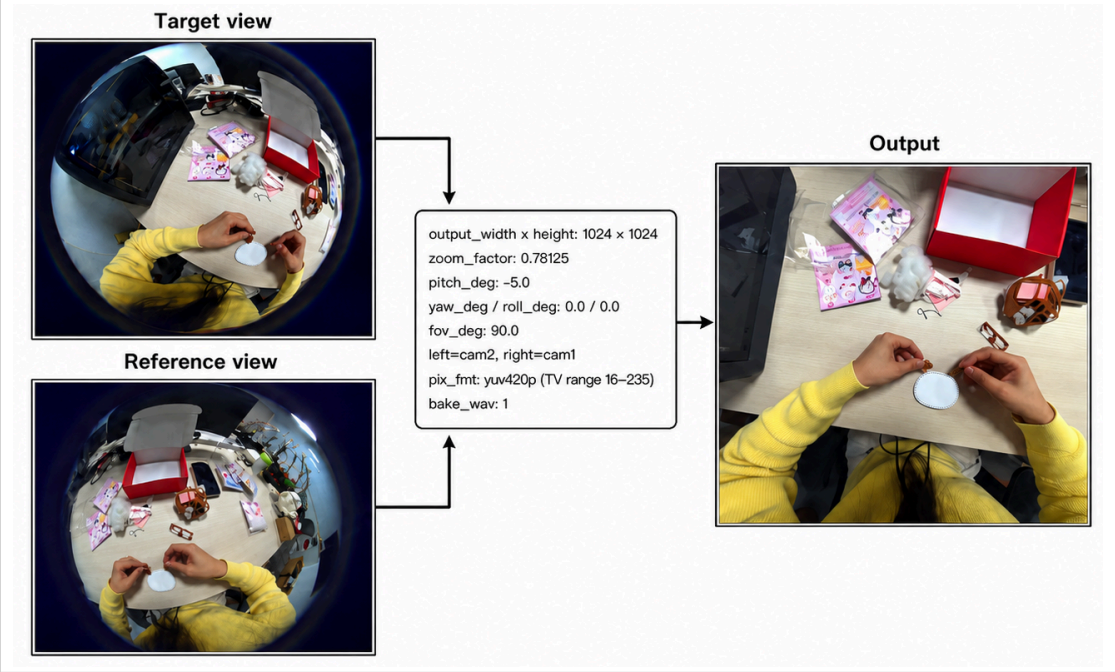


Figure 6 | **Rectification setting example.** Given a target view and a reference view from the same recording, the rectification process applies fixed camera parameters to produce a spatially normalized egocentric frame. The output reduces fisheye distortion while preserving the hand–object interaction 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 of narration is to capture what tool is being used, what alternative tool may be relevant, what target is being acted on, and why the selected tool is appropriate in the current context.

The interface allows participants or data collectors to watch rectified egocentric videos, mark timestamps, and write tool-centric narrations. The template encourages narrations to specify the selected tool, a relevant alternative when applicable, the target object, the action, and the physical rationale. Language tags indicate the original narration language, while the displayed text is the corrected English version used as corrected English narration annotations. These narrations are collected as corpus annotations only; they are not provided to EgoTools-Bench annotators as answer evidence or to models during benchmark evaluation.

Tool-centric narration style. Participants are encouraged to provide free-form but tool-centric narrations. The following template is used as a reference rather than a mandatory fixed form: [Tool-centric Narration Template] `I am using [the] [adj./property] [tool] [location/identity cue]`
`rather than [the] [adj./property] [tool] [location/identity cue]`
`to [action] on/for [target] because [reason].`

Narration target. Narrations focus on tool-use moments that involve meaningful decisions, such as choosing a tool, switching tools, adapting to object state, preparing for a later step, or using a tool in a physically appropriate way. Participants are not required to narrate every visible action.

Narration language and signal. Each narration describes the selected tool, the target object, the reason for the tool choice, and a relevant alternative when applicable. Participants may speak in Korean, English, or Chinese; the speech is transcribed, translated into English, and corrected when necessary. The corrected English text is paired with the corresponding video segment and used as the primary narration annotation signal, while the original narration audio is preserved as a dataset asset for future omni-modal extensions and release, subject to the written release agreement and privacy review.

Grounding-oriented annotations. Narrations are associated with focal tools. 2D grounding is annotated for selected instances and will be progressively released as annotations are finalized. These annotations support focal-tool localization, trajectory construction, and future grounded video-language training.

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 comprises approximately 100 hours of narrated egocentric video collected across seven tool-use domains: kitchen, classroom, research laboratory, repair workshop, craft, office, and household. From this corpus, we select approximately 40.34 hours of high-quality, tool-dense videos for benchmark construction.

Distribution statistics. EgoTools-Bench contains **1,000** 8-way multiple-choice question–answer pairs, including **900** human-crafted questions and **100** human-verified pipeline-generated spatial reasoning questions. 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-use reasoning while maintaining broad coverage of perceptual grounding, procedural understanding, and spatial hand–tool–object relations. All benchmark clips and QA annotations are reserved exclusively for evaluation.

Table 6 | **Summary of EgoTools dataset and benchmark statistics.**

Statistic	Value
Total EgoTools video	100 hours
Benchmark construction subset	40.34 hours
Benchmark QA pairs	1,000
Human-crafted QAs	900
Pipeline-generated QAs	100

E.2. Release and Privacy Considerations

The EgoTools release is scoped to research use and includes only materials cleared by the applicable release authorization, anonymization requirements, and privacy review. The release

focuses on rectified egocentric videos rather than raw fisheye recordings, together with the annotations and metadata needed to support benchmark evaluation and dataset documentation.

Consent and release scope. Collected materials are used and released only within the applicable participant, contributor, and institutional authorization scope. Release components are governed by anonymization procedures and privacy-review requirements. Materials that do not satisfy these conditions are excluded from the public release.

Privacy safeguards. We remove or anonymize names, contact information, institutional identifiers, and other personally identifiable information when identified. Sensitive, private, or accidental content is reviewed before release, and any content deemed unsuitable for public distribution is excluded.

Release components. The release package includes the following cleared components:

- Rectified egocentric video clips;
- Corrected English text narrations;
- Original narration audio, when covered by release authorization and privacy review;
- Benchmark QA pairs;
- Clip metadata, timestamps, and split information;
- Selected finalized grounding-oriented focal-tool annotations;
- 3D trajectory visualizations for selected clips as part of the release materials.

Excluded components. Raw fisheye videos will not be publicly released.

Access conditions. EgoTools is released for research use. Privacy-sensitive components are handled according to the applicable release authorization, anonymization requirements, and privacy-review outcomes.

License. The dataset is distributed under a research-use license that specifies permitted uses, redistribution constraints, and usage restrictions.

Limitations. EgoTools is designed to support the study of tool-use reasoning across multiple real-world environments; however, it does not exhaustively cover all tool-use domains. Grounding annotations and audio components are released only for examples that have passed the corresponding annotation-finalization, release-authorization, and privacy-review checks.

Table 7 | **Anonymized contributor profiles.** IDs are randomly assigned and do not correspond to names, institutional identifiers, or public profiles. Role abbreviations: P = participant, N = narrator, QA = QA annotator, R = QA reviewer, C = data curator, G = grounding annotator, T = trajectory annotator. Degree abbreviations: UG = undergraduate, MS = master's student. Field abbreviations: CS = computer science, EE = electrical engineering.

Anon. ID	Role(s)	Degree	Field	Relevant Expertise
C01	P / N / QA / C	PhD	CS	Kitchen tool use
C02	T / G	Postdoc	CS	3D trajectory; visual grounding
C03	P / N / QA / C	UG	EE	Device manipulation
C04	P / N	MS	CS	Textile work
C05	P / N / QA	UG	Film	Video capture; visual storytelling
C06	P / N / QA	UG	CS	Egocentric data collection; tool-use reasoning; benchmark design
C07	P / N / QA	UG	EE	Electronics handling; egocentric video collection
C08	P / N	UG	EE	Electronics handling; egocentric video collection
C09	P / N	UG	EE	Egocentric video understanding; tool-use reasoning
C10	P / N / QA	MS	CS	Textile work
C11	R	MS	CS	Multimodal evaluation; benchmark quality control
C12	R	MS	CS	Multimodal evaluation; benchmark quality control
C13	T	MS	CS	3D trajectory; visual grounding
C14	P / N	UG	EE	Egocentric video understanding; tool-use reasoning
C15	P / N / QA / C	UG	CS	Cooking
C16	P / N	PhD	Biology	Laboratory procedures
C17	QA	UG	CS	Egocentric video understanding; tool-use reasoning
C18	QA	UG	CS	Egocentric video understanding; tool-use reasoning
C19	P / N	PhD	Biology	Laboratory procedures
C20	QA	UG	EE	Egocentric video collection
C21	P / N	PhD	Chemistry	Chemistry laboratory procedures
C22	P / N	UG	EE	Circuit assembly
C23	P / N	PhD	CS	Cooking

F. Quality Control and Normalization

F.1. Fix-First Quality Control

Raw human annotations are processed under a fix-first quality-control policy. Annotator effort is treated as a limited resource, and items are repaired whenever a deterministic or model-assisted edit can preserve the original intent.

Structural audit. Each item is checked using a 22-flag boolean audit for defects such as placeholders, missing fields, malformed options, duplicate entries, first-person leakage, personally identifiable information, answer-option mismatch, full-video clip misuse, and distractor anomalies.

Filtering policy. Content-empty submissions and verbatim duplicates are removed. Repairable defects are corrected through deterministic rules or model-assisted minimal edits. Low-confidence cases are routed to human review rather than automatically accepted.

Model-assisted repair. We use Gemini-2.5-flash-lite at temperature 0 for model-assisted repair. The prompt asks the model to perform the smallest meaning-preserving edit possible. Examples include replacing personally identifying names with role tokens, rewriting first-person references into third-person form, correcting surface grammar, and reconstructing truncated answers only when context is unambiguous.

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

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

Anti-shortcut checks. Each normalized option set is passed through deterministic anti-shortcut checks before finalization. These 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, excessive question–answer lexical overlap, answer-position bias, and answer-length bias. Failed items are routed to bounded retry or human review, where previous violations are used as feedback for rewriting the options.

Length-bias check. For each item, we compare the character length of the correct answer against the distractor-only length distribution. Let L_{ans} denote the correct-answer length, and let μ_{dist} and σ_{dist} denote the mean and standard deviation of distractor lengths. We compute

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

Items outside the accepted $\pm 1.5\sigma$ band are flagged for repair. Cached eight-option snapshots from prior evaluation runs are subjected to the same checks rather than accepted blindly, closing a bypass that had allowed a corpus-level length skew of $\bar{z} \approx +2.21\sigma$ to leak into the benchmark. With the check enforced, the surviving set has $\bar{z} \approx -0.15\sigma$, with no item outside the accepted $\pm 1.5\sigma$ range.

Option shuffling. Final option order is shuffled using a deterministic seed derived from the QA identifier. This reduces positional priors while keeping evaluation reproducible. Since the released benchmark is normalized to eight choices, the main benchmark tables report standard multiple-choice accuracy under a consistent chance baseline.

G. Additional Related Work

Broader egocentric video reasoning. Egocentric video benchmarks have expanded from action and object understanding toward memory, planning, temporal localization, and assistance. QAEgo4D and GroundVQA emphasize episodic-memory QA and grounded temporal QA in long egocentric videos [5, 11]. EgoTaskQA evaluates task-oriented questions about dependencies, effects, goals, and beliefs [23]; EgoSchema and EgoTempo focus on long-context and temporal reasoning [32, 37]; and EgoPlan-Bench evaluates planning from first-person videos [8]. Recent benchmarks further study first-person thinking, life-log assistance, real-time aid, expert-guided assistance, situated awareness, multi-agent egocentric QA, intent prediction, personalized grounding, and AR/life-logging environments [9, 14, 24, 26, 35, 41, 47, 52, 55, 60]. These resources broaden evaluation around memory, temporal context, planning, and user assistance, but they generally do not isolate the physical question of why a particular tool is appropriate, how it can be replaced, or how its use should change as the target object changes state.

Instructional, embodied, and robotic interaction benchmarks. Instructional-video datasets provide useful comparisons because they often contain tools, ingredients, techniques, and procedural alternatives. VidDetours retrieves detours from how-to videos, and StepDiff compares

differences between instructional clips [1, 33]. However, their main focus is procedural comparison or retrieval rather than first-person physical tool-use reasoning grounded in an ongoing 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 for robotic demonstrations [45], and RoboCasa365 studies simulated kitchen manipulation across many tasks and environments [34]. These benchmarks are useful scope contrasts, but they do not provide the same combination of natural egocentric human video, tool-centric narration, narration-linked tool grounding, and explicit QA over tool choice, substitution, temporal manipulation, and state-dependent adaptation.

Affordance grounding and physical tool understanding. Robotics and embodied AI have long treated tool use as a problem of affordance, grasping, motion, and task execution. Task-oriented grasping studies how tools should be grasped for tasks such as sweeping and hammering [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 masks, and physically grounded policies [6, 15, 27, 39, 54]. Recent multimodal benchmarks and methods further test whether MLLMs and VLMs can recognize affordances, infer physical constraints, and ground tool-related interactions [22, 38, 57, 59, 56]. These works reveal persistent weaknesses in tool functions, constraints, and multi-step interactions. Nevertheless, most evaluations are built from static images, synthetic or 3D scenes, short interactions, or constrained robotic setups. EgoTools-Bench is complementary: it evaluates physical tool-use reasoning from natural first-person videos, where the model must interpret real human tool selection, substitution, 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 benchmarks ask models to infer user needs and intentions from egocentric observations [46]. These settings motivate first-person assistance, but they do not specifically target tool-centered physical reasoning. EgoTools-Bench therefore complements wearable-assistant work by focusing on questions that require understanding the functional role of tools, feasible substitutes, ordered manipulation, and adaptation to changing object states.