EGOACTIONBENCH: BENCHMARKING DENSE ACTION CAPTIONING FOR EGOCENTRIC HUMAN VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from human data offers a scalable path toward embodied intelligence, yet it requires models to accurately understand fine-grained human actions from egocentric observations. However, existing benchmarks remain limited in action atomicity, temporal precision, and hand–action association. We introduce EGOACTIONBENCH, a benchmark that represents continuous human manipulation as temporally localized atomic actions associated with the left hand, right hand, or both hands. EGOACTIONBENCH contains 1,407 episodes with 40,384 human-refined action annotations. To evaluate such dense predictions, we propose a progressive evaluation protocol that first retrieves temporally eligible candidates, selects a prediction by semantic agreement, and then scores its temporal extent and hand association, jointly capturing what happens, when it happens, and which hand performs it. We benchmark 19 vision-language models on EGOACTIONBENCH and further perform supervised fine-tuning on two backbones, Qwen3-VL-4B and LLaVA-OneVision-2-8B, using automatically annotated training episodes. Fine-tuning improves EgoActionScore by 10.4 and 20.0 points on the two backbones, respectively, and both adapted models outperform all open-source baselines. Our experiments reveal substantial limitations of current models in action atomicity, temporal precision, and hand–action association, highlighting the gap between general video understanding and the action understanding required for embodied intelligence.

1 INTRODUCTION

Learning from human data offers a scalable path toward embodied intelligence (Nair et al., 2023; Kareer et al., 2025; Punamiya et al., 2026; Cao et al., 2026; Wang et al., 2026b; Zheng et al., 2026; Li et al., 2026c; Deng & Zhou, 2026; Wang et al., 2026c; Gao et al., 2026). Egocentric video is particularly useful in this context, as it records human interactions with objects from the actor's viewpoint and provides rich observations of everyday manipulation (Damen et al., 2022; Grauman et al., 2022; 2024; Perrett et al., 2025; Hoque et al., 2026; Yang et al., 2025; Li et al., 2026b; Paul et al., 2026; Feng et al., 2026). However, extracting useful action supervision from such videos requires more than recognizing the overall activity. Models must recover the individual actions within continuous manipulation, determine when each action occurs, and identify which hand performs it.

Existing video understanding benchmarks capture only parts of this problem. Action recognition typically maps predefined video segments to action categories (Kay et al., 2017; Goyal et al., 2017; Damen et al., 2022), while video captioning often describes an entire clip or a relatively coarse event (Xu et al., 2016; Krishna et al., 2017; Zhou et al., 2018a; Ohkawa et al., 2025). Temporal action localization requires models to recover action boundaries, but is commonly formulated over predefined action vocabularies or events at a coarser granularity (Caba Heilbron et al., 2015; Zhang et al., 2022a; Wang et al., 2023). Recent egocentric benchmarks have moved closer to understanding everyday human activities and hand–object interactions (Darkhalil et al., 2022; Grauman et al., 2024; Perrett et al., 2025; Tateno et al., 2026). However, these formulations do not jointly address three requirements for recovering fine-grained human manipulation: **action atomicity**, **temporal precision**, and **hand-action association**.

Action atomicity requires decomposing continuous manipulation into individual actions rather than summarizing a sequence with a single activity label. For example, an interval described as "stirring"

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

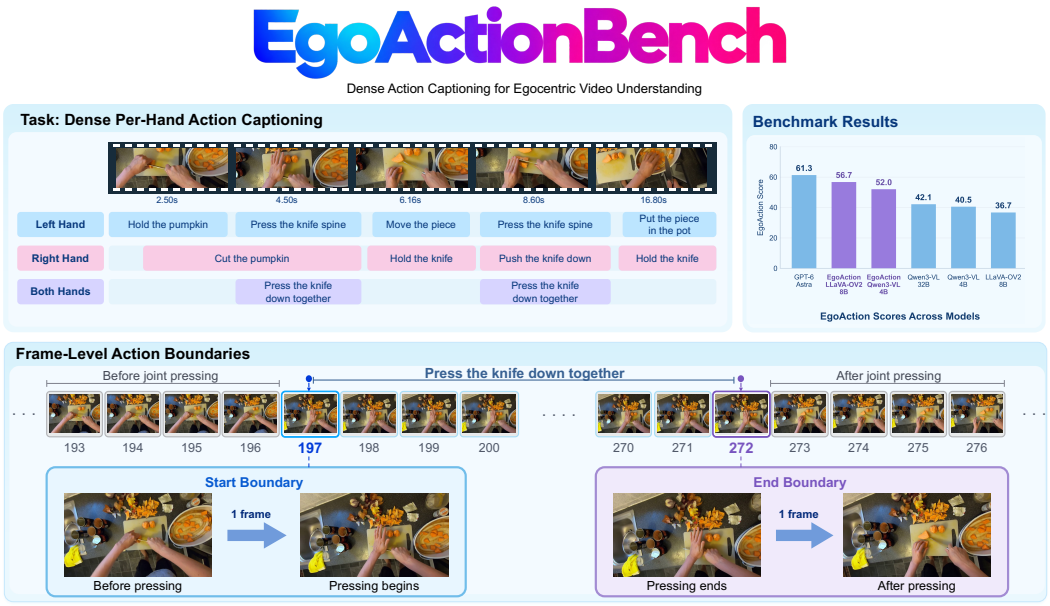

Figure 1: **Overview of EgoAction.** Supervised adaptation trains EgoAction models to emit dense timestamped hand actions, while EgoActionBench evaluates whether the output names the right action, at the right time, on the right hand track.

may include picking up a spoon, stirring with it, and putting it down. Each action captures a distinct interaction with the object. A fine-grained description should preserve these actions and specify the objects involved, distinguishing, for example, stirring with a spoon from stirring with chopsticks.

Temporal precision requires identifying the start and end of each action, including actions that last less than a second. For example, a person washing dishes may briefly move the faucet handle to turn on the water. Missing this short action leaves a gap in the recovered sequence, even if the surrounding activity is correctly recognized as "washing dishes." Recovering such actions requires sufficiently fine temporal resolution to distinguish them from adjacent actions and localize their boundaries.

Hand-action association requires assigning each action to the left hand, right hand, or both hands while preserving their temporal relationships. The two hands can perform different actions over different intervals: the left hand may hold a bowl throughout, while the right hand picks up a spoon, stirs, and puts it down. Describing the interval as "holding a bowl and stirring" obscures which hand performs each action and how the actions overlap. A fine-grained representation should retain the sustained left-hand action alongside the sequence of shorter right-hand actions.

To study this capability, we introduce EGOACTIONBENCH, a benchmark for fine-grained egocentric action understanding. EGOACTIONBENCH represents continuous human manipulation as temporally localized atomic actions associated with the left hand, right hand, or both hands. It contains 1,407 episodes with 40,384 human-refined action annotations. Rather than providing models with predefined temporal segments to classify or describe, the benchmark requires them to recover the action structure directly from continuous egocentric video. Its annotations capture short-lived actions and independent hand behaviors with fine temporal granularity, enabling systematic evaluation beyond coarse activity descriptions.

Evaluating such predictions also requires careful treatment of temporal segmentation. Predicted and reference actions can differ in both their temporal boundaries and semantic content. A single prediction may span multiple reference actions, while several predictions may fragment one reference action. Existing evaluation protocols based on merging or strict one-to-one matching can therefore confound segmentation differences with semantic errors (Fujita et al., 2020; Yao et al., 2026). We propose a progressive evaluation protocol that first retrieves temporally eligible predictions, selects

one by semantic agreement, and then scores its temporal extent and hand association. This ordering ensures that semantic evaluation is performed over temporally corresponding actions while retaining the ability to diagnose localization errors.

We benchmark 19 vision-language models on EGOACTIONBENCH, spanning proprietary models, open-source general-purpose VLMs, video captioning models, and egocentric video understanding models, and analyze their ability to recover fine-grained actions from continuous egocentric video. Our experiments reveal substantial limitations in current models, including missed short actions, inaccurate temporal boundaries, merged action sequences, and incorrect associations between actions and hands. We further perform supervised fine-tuning on two backbones, Qwen3-VL-4B and LLaVA-OneVision-2-8B, using automatically annotated training episodes, studying whether supervision at the target granularity can improve action understanding without modifying the underlying model architectures. Fine-tuning improves EgoActionScore by 10.4 and 20.0 points on the two backbones, respectively, and both adapted models outperform all open-source baselines. These results establish baselines for EGOACTIONBENCH and highlight the gap between general video understanding and the detailed action understanding needed to learn effectively from human data.

Our main contributions are:

- We introduce EGOACTIONBENCH, a benchmark containing 1,407 egocentric episodes and 40,384 human-refined action annotations, representing continuous human manipulation as temporally localized atomic actions associated with the left hand, right hand, or both hands.
- We propose a progressive evaluation protocol that first retrieves temporally eligible predictions and then uses semantic agreement to select a shared correspondence for all three scores, enabling fine-grained evaluation when predicted and reference actions differ in temporal segmentation.
- We benchmark 19 vision-language models and perform supervised fine-tuning on Qwen3-VL-4B and LLaVA-OneVision-2-8B using automatically annotated training data, revealing persistent challenges in recovering fine-grained action structure from continuous egocentric video.

2 RELATED WORK

2.1 EGOCENTRIC VIDEO UNDERSTANDING

Egocentric action understanding has progressed from transferable video–language representations, to interaction-aware modeling, and finally to language-grounded descriptions of ongoing activity. This progression is particularly relevant to embodied systems, where first-person observations connect perception to the hand and object being manipulated. Large-scale resources and narration-based pretraining established the foundation for egocentric learning (Damen et al., 2022; Grauman et al., 2022; 2024; Lin et al., 2022; Zhao et al., 2023; Pramanick et al., 2023). Subsequent work made hand–object relations, bimanual structure, motion, and spatial cues more explicit through dedicated supervision and interaction-aware models (Darkhalil et al., 2022; Zhang et al., 2022b; Kwon et al., 2021; Liu et al., 2024; Pei et al., 2025a; Xu et al., 2025). More recent video-language models and cross-view methods use language for long-video description, question answering, and knowledge transfer (Ren et al., 2024; Ye et al., 2025; Bai et al., 2025a; Zhang et al., 2026; Li et al., 2026a).

2.2 VIDEO CAPTIONING BENCHMARKS

Video captioning benchmarks mainly differ in the granularity of the described event, the output required from the model, and the evidence used for evaluation. Early third-person benchmarks established dense captioning for general events and procedural steps (Krishna et al., 2017; Zhou et al., 2018a;b). Egocentric extensions vary between single-activity captioning, hierarchical narration, and localized procedural description (Nakamura et al., 2021; Islam et al., 2024; Ohkawa et al., 2025; Shi et al., 2025). For embodied manipulation, WGO-Bench adds instruction-conditioned timestamped subtask annotation, but does not require an explicit acting-hand track (Macrodata Labs, 2026a;b). HanDyVQA is complementary: it evaluates hand–object reasoning through question-conditioned QA rather than open-ended caption generation and autonomous temporal discovery (Tateno et al., 2026).

Table 1: **Comparison with related datasets and benchmarks.** Temporal localization and caption generation denote evaluated tasks; atomic action evaluation denotes field-level caption scoring. Ego4D and Ego-Exo4D refer to narrations; EgoVerse refers to EgoVerse-I. †25% human egocentric episodes; the remainder are robot videos.

Dataset / Benchmark	Scene	Target unit	Ego.	Temp. loc.	Caption gen.	Hand-specific Captions	Atomic act. eval.
ActivityNet Captions (Krishna et al., 2017)	Real-world	Events	✗	✓	✓	✗	✗
EPIC-KITCHENS-100 (Damen et al., 2022)	Kitchen	Actions	✓	✓	✗	✗	✗
Ego4D (Grauman et al., 2022)	Real-world	Events	✓	✗	✗	✗	✗
Ego-Exo4D (Grauman et al., 2024)	Real-world	Events	✓	✗	✗	✗	✗
EgoYC2 (Ohkawa et al., 2025)	Kitchen	Procedural steps	✓	✓	✓	✗	✗
HD-EPIC (Perrett et al., 2025)	Kitchen	Actions	✓	✗	✗	✓	✗
VidCapBench (Chen et al., 2025)	Real-world	Clips	✗	✗	✓	✗	✗
HanDyVQA (Tateno et al., 2026)	Real-world	Actions	✓	✗	✗	✗	✗
WGO-Bench (Macrodata Labs, 2026a)	Manipulation	Procedural steps	25%†	✓	✓	✗	✗
OmniDCBench (Yao et al., 2026)	Real-world	Scenes	✗	✓	✓	✗	✗
CAPEval (Liu et al., 2026)	Diverse	Images	✗	✗	✓	✗	✗
EgoVerse (Punamiya et al., 2026)	Manipulation	Procedural steps	✓	✗	✗	✓	✗
EgoActionBench (ours)	**Real-world**	**Atomic actions**	✓	✓	✓	✓	✓

On the other hand, evaluation protocols range from caption similarity and temporal localization to factuality checks and question-answer accuracy (Liu & Wan, 2023; Chen et al., 2025; Yao et al., 2026). These settings cover important aspects of event description and interaction reasoning, but do not jointly evaluate the content, interval, and acting hand of the same prediction. EgoActionBench instead takes a complete egocentric clip without an action query or reference boundary and requires free-form captions on Left, Right, and Both tracks.

3 EgoActionBench

EgoActionBench evaluates dense hand-action captioning in egocentric videos. We first define the task and its output format, then describe data collection and annotation, and finally present a GT-anchored evaluation protocol that separately measures semantic content, temporal localization, and hand assignment.

3.1 Task Definition

We introduce the task of *dense per-hand action captioning* for egocentric video understanding. Given a complete egocentric video clip, a model must temporally localize and describe the camera wearer's visible atomic hand actions, with an explicit hand assignment for each action. Each prediction is represented as

$$p_i = (\hat{s}_i, \hat{e}_i, \hat{h}_i, w_i), \tag{1}$$

where $[\hat{s}_i, \hat{e}_i)$ is the predicted temporal interval, $\hat{h}_i \in \{L, R, B\}$ denotes the Left, Right, or Both track, and w_i is a free-text action caption. No action queries or reference boundaries are provided.

Each caption describes one atomic action: a continuous visible operation delimited by meaningful changes in the operation, manipulated object, contact relation, or dominant motion. The Left and Right tracks describe actions of the corresponding anatomical hands. The Both track captures coordinated manipulation of a shared object, rather than simultaneous independent actions. Both captions describe the collaboration, while any annotated single-hand roles remain separate records on Left and Right.

3.2 Data Collection

Data sources and selection. Episodes are curated from Ego4D (Grauman et al., 2022), Ego-Life (Yang et al., 2025), and EPIC-KITCHENS-100 (Damen et al., 2022). Each episode captures a coherent task or subtask performed by the camera wearer and is screened for suitability for dense

Table 2: **Dataset Statistics of EgoActionBench.** Video clips, total video duration, and human-refined action annotations by source. Left, Right, and Both report annotation counts for the corresponding hand tracks.

Source	Clips	Hours	Actions	Left	Right	Both
Ego4D	451	3.130	16,432	6,711	8,156	1,565
EgoLife	485	2.676	13,147	5,847	6,037	1,263
EPIC-KITCHENS-100	471	2.176	10,805	4,500	5,191	1,114
Total	1,407	7.982	40,384	17,058	19,384	3,942

hand-object annotation. Sufficient surrounding context is retained to distinguish the target operations from neighboring activities. Episodes with severe occlusion, unverifiable actions, or unrelated task transitions are discarded. Source-video identifiers and original temporal ranges are retained to trace each episode and support separation from the adaptation data.

Dataset statistics. The collection contains 1,407 episodes with 40,384 action captions across the Left, Right, and Both tracks (table 2). It spans 7.982 hours, with a mean episode duration of 20.42 seconds and a median action duration of 0.63 seconds. Non-action entries, including invisibility, idle intervals, and the absence of bimanual collaboration, are excluded from the action counts.

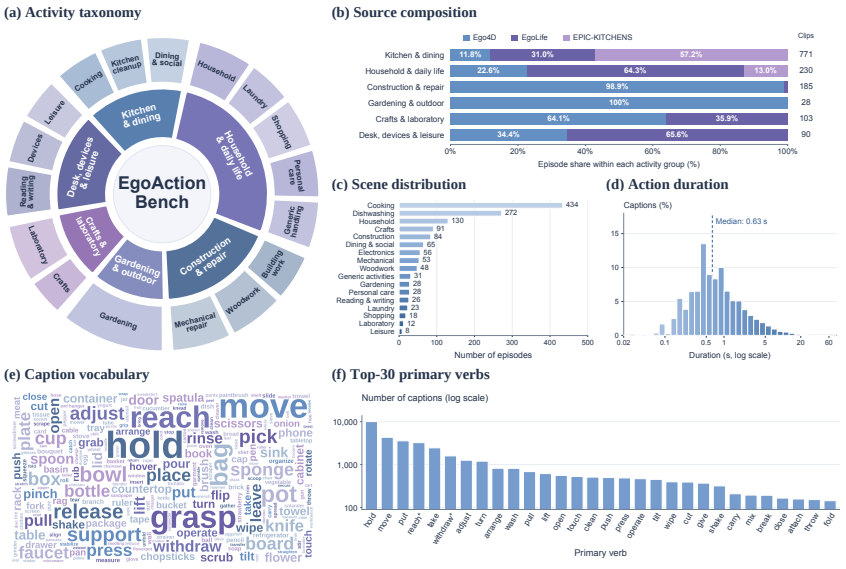

Figure 2: **Dataset Composition and Action Statistics.** (a) Activity taxonomy with six groups and 17 scenes. (b) Source composition within each activity group. (c) Episode counts by scene. (d) Action-duration distribution on a logarithmic time axis, with a median of 0.63 s. (e) Caption vocabulary. (f) The 30 most frequent primary verbs, with counts on a logarithmic scale. Statistics cover 1,407 episodes and 40,384 action captions; sector angles do not encode counts.

3.3 Data Annotation Pipeline

We annotate the selected episodes through a five-stage human-in-the-loop pipeline (fig. 3). First, CaRe-Ego (Su et al., 2025) provides temporally smoothed hand masks to highlight the left and right hands. A VLM then generates sliding-window captions for the Left, Right, and Both tracks, merging consecutive fragments of the same continuous action into draft annotations. Human annotators correct action descriptions, hand assignments, and temporal boundaries against the original video. Agentic quality control subsequently inspects local visual evidence to propose further revisions, guided by hand-object contact, hand-motion phases, and object-state changes. Finally, human reviewers verify the annotations; episodes requiring further correction return for annotation and QC. Only human-verified annotations enter the benchmark. Detailed annotation and quality-control procedures are provided in section B.

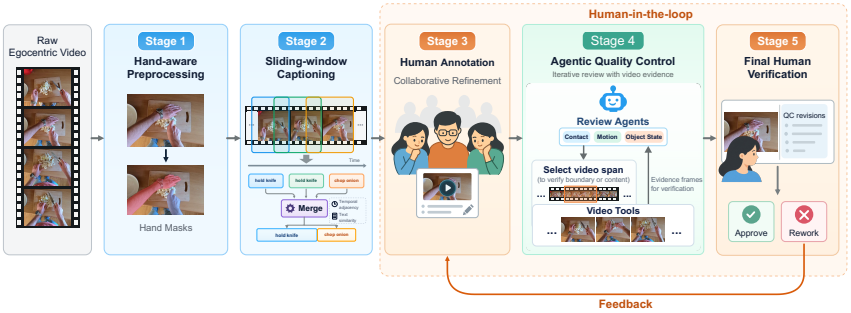

Figure 3: **Human-in-the-Loop Construction of EgoActionBench.** Hand-aware preprocessing and sliding-window captioning generate draft per-hand annotations. Human annotation, agentic quality control, and final human verification refine these drafts against the original video. Episodes requiring further correction return to human annotation and undergo QC again.

3.4 Evaluation Suite

We evaluate **what** action is described, **when** it occurs, and **which hand** performs it through Semantic, Temporal, and Hand scores. Our evaluation is explicitly **ground-truth-based**: each ground-truth action selects at most one prediction, and all three scores use that same prediction. As illustrated in fig. 4, we first retrieve temporally eligible predictions from compatible hand tracks, then select the candidate with the highest semantic agreement. A language-model judge assesses semantic agreement across four fields: Verb, Objects, Tools, and Motion. The selected prediction is scored using this semantic assessment, temporal IoU, and hand-label agreement, with its caption and interval retained in full. Ground-truth actions without eligible predictions receive zero on all three dimensions. Each component is averaged over ground-truth actions on a 0–100 scale, and their arithmetic mean defines **EgoActionScore**. Full matching rules, scoring definitions, and metric interpretation are provided in section C.

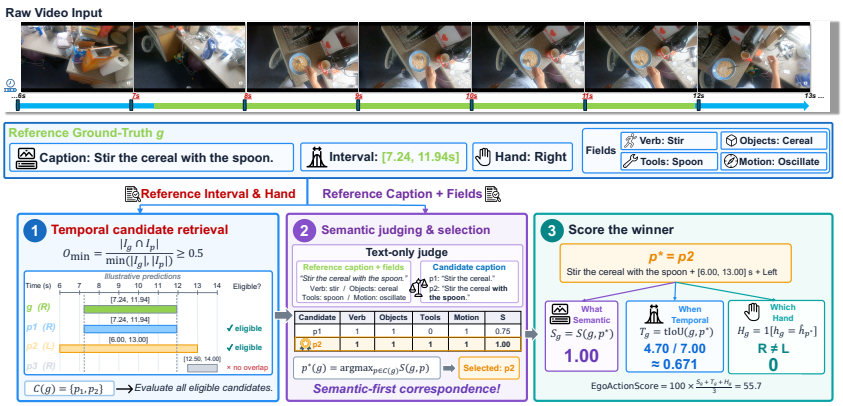

Figure 4: **EgoAction evaluation pipeline.** For each ground-truth action, temporal overlap and hand-track compatibility determine the eligible candidates. Field-level semantic agreement selects one prediction, whose original caption, interval, and hand label are used to compute Semantic, Temporal, and Hand scores. Each component is averaged separately over all ground-truth actions.

4 Learning to Caption Egocentric Actions

We study whether supervision at the granularity defined in section 3 improves fine-grained egocentric action understanding. Given a video, the model must recover individual actions, their temporal boundaries, and the hands that perform them. We learn this mapping by supervised fine-tuning on examples that express these quantities jointly.

Training data. The training set contains 5,354 egocentric video episodes drawn from Ego4D, EPIC-KITCHENS, and EgoLife, including 5,146 training examples and 208 validation examples. Each example pairs a video with timestamped captions organized into Left, Right, and Both tracks. The targets jointly supervise action descriptions, temporal intervals, and hand assignments. Unlike the human-refined annotations used to evaluate EgoActionBench, these training labels are generated and refined automatically. Source identifiers are retained to separate the adaptation data from the benchmark evaluation set.

Supervised adaptation. We perform supervised fine-tuning on two pretrained backbones: Qwen3-VL-4B-Instruct (Bai et al., 2025a) and LLaVA-OneVision-2-8B (An et al., 2026). Qwen3-VL serves as a representative general-purpose model for image and video understanding. LLaVA-OneVision-2 provides a complementary backbone through its native-resolution OneVision-Encoder, unified spatiotemporal modeling, and emphasis on temporal and spatial grounding. This pairing allows us to examine whether dense action supervision benefits different pretrained models, including one already oriented toward fine-grained video perception. Given a video and a task instruction, each model is trained to generate timestamped action captions across the Left, Right, and Both tracks. Action descriptions, temporal intervals, and hand assignments are jointly supervised through next-token prediction, with the loss applied only to the target response. The original architectures are retained, with the visual encoders frozen and the language models and multimodal connectors updated. We refer to the adapted models as EgoAction, with each variant identified by its backbone.

5 Experiments

5.1 Experimental Setup

We evaluate 19 baseline models on EgoActionBench, spanning four categories: proprietary models, open-source general-purpose VLMs, video captioning models, and egocentric video understanding models. This comparison examines how general multimodal capability and task specialization translate to dense action captioning. The EgoAction models introduced in Section 4 are evaluated alongside their Qwen3-VL-4B and LLaVA-OneVision-2-8B backbones. Models receive video inputs through their respective input interfaces and are prompted to follow the task specification in Section 3.1. Evaluation follows the suite in Section 3.4, with Qwen3-32B serving as the semantic judge. We report EgoActionScore together with Temporal, Hand, Semantic, and the four semantic-field scores.

Human baseline. To establish a human baseline, 3 annotators independently annotate evaluation episodes following the task definition in Section 3.1. They identify action boundaries, assign hand tracks, and write action captions from the original videos, without access to the reference annotations or model predictions. Their annotations are evaluated against the fixed benchmark references using the same matching procedure, semantic judge, and scoring rules as the models. The resulting scores are reported as Human Expert in Table 3.

5.2 Main Benchmark Results

Table 3 presents the main results on EgoActionBench. Both EgoAction models outperform all open-source baselines: the models based on LLaVA-OneVision-2-8B and Qwen3-VL-4B reach EgoActionScores of 56.7 and 49.7, respectively, compared with 43.9 for the strongest open-source baseline, CapRL-Video-4B. These results show that task-aligned adaptation can make compact open-source backbones competitive with proprietary models on dense action captioning.

General capability and task specialization. Among general-purpose open-source VLMs, Qwen3-VL-32B achieves the highest EgoActionScore of 42.1. Model size does not produce a monotonic ranking: Qwen3-VL-4B scores 39.3, compared with 36.2 for its 8B counterpart and 38.1 for Qwen2.5-VL-72B. Specialized models also exhibit substantial variation. CapRL-Video-4B achieves 43.9, whereas the egocentric models EgoThinker-7B and EgoGPT-7B score 43.7 and 27.2, respectively. Captioning or egocentric specialization therefore does not consistently translate into stronger performance on the combined demands of action content, temporal localization, and hand assignment.

Table 3: **Dense action captioning on EgoActionBench.** All scores range from 0 to 100; higher is better. Open-source VLMs are ordered by nominal backbone size; parentheses identify the EgoAction backbone. Bold marks the best model score in each column.

Model	EgoAction Score	Semantic					Temporal	Hand
		Verb	Objects	Tools	Motion	Overall		
Human Baseline								
Human Expert	92.5	96.1	95.7	95.7	75.8	89.4	92.2	96.0
Proprietary models								
GPT-5.6 Sol (OpenAI, 2026a)	54.2	51.4	60.7	77.0	41.7	51.5	36.5	74.6
GPT-6 Astra (OpenAI, 2026b)	**61.3**	**63.3**	**69.8**	**80.4**	**48.9**	**60.8**	**44.6**	78.4
Gemini 3.1 Pro Preview (Google DeepMind, 2026)	41.0	35.9	44.3	64.9	30.8	37.3	24.4	61.2
Open-source VLMs								
Qwen2.5-VL-72B-Instruct (Bai et al., 2025b)	38.1	26.8	28.7	43.8	16.9	24.3	14.1	75.8
Qwen3-VL-32B-Instruct (Bai et al., 2025a)	42.1	32.1	37.4	58.8	29.9	33.3	17.1	75.9
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	36.2	26.2	27.5	42.4	18.1	24.1	16.8	67.5
InternVL3.5-8B (Wang et al., 2025)	31.9	20.8	20.0	31.2	14.4	18.5	15.5	61.7
MiniCPM-V-4.5-8B (OpenBMB, 2025)	38.0	23.5	23.4	29.3	17.5	21.5	18.6	73.8
Molmo2-8B (Clark et al., 2026)	24.3	18.3	19.0	31.3	13.0	17.0	12.5	43.5
LLaVA-OV2-8B (An et al., 2026)	36.7	26.6	32.7	47.5	24.4	28.1	14.4	67.8
VideoLLaMA3-7B (Zhang et al., 2025)	25.3	17.9	25.5	39.3	16.6	20.1	7.6	48.0
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	39.3	26.9	29.9	40.3	18.7	25.4	19.3	73.1
Open-source video captioning models								
CapRL-Video-4B (Yang et al., 2026)	43.9	29.1	39.3	61.1	27.9	32.4	25.6	73.7
OmniCaptioner-IF-3B (Wang et al., 2026a)	38.7	19.0	22.3	28.9	19.6	20.3	13.4	82.4
OmniCaptioner-IF-7B (Wang et al., 2026a)	40.1	25.8	26.6	40.9	24.8	25.9	19.3	75.2
video-SALMONN-2+-3B (Tang et al., 2025)	22.2	8.4	14.1	29.0	9.1	10.7	7.9	47.9
Open-source egocentric video understanding models								
EgoThinker-7B (Pei et al., 2025b)	43.7	20.7	19.6	13.6	19.7	20.0	31.5	79.5
EgoGPT-7B (Yang et al., 2025)	27.2	19.5	18.7	22.9	20.5	19.6	8.8	53.0
RynnBrain-4B (Dang et al., 2026)	25.3	13.2	15.9	15.6	14.6	14.6	19.4	41.8
EgoAction (Ours)								
EgoAction (LLaVA-OV2-8B)	56.7	43.4	47.1	55.9	40.1	43.7	40.1	**86.3**
EgoAction (Qwen3-VL-4B)	49.7	32.9	32.9	42.8	25.2	30.5	36.1	82.4

Supervised adaptation across backbones. Adaptation improves EgoActionScore from 39.3 to 49.7 for Qwen3-VL-4B and from 36.7 to 56.7 for LLaVA-OneVision-2-8B. Both backbones improve in all three main dimensions. Temporal increases from 19.3 to 36.1 and from 14.4 to 40.1, respectively; Hand and Semantic also improve for both models. Both adapted models also improve across all four semantic fields. The aggregate gains thus reflect improvements in multiple aspects of action recovery, particularly temporal localization, with backbone-dependent gains in hand assignment.

Temporal precision and motion remain challenging. The component scores reveal a persistent gap between describing an action and recovering its temporal extent. GPT-5.6 Sol obtains 51.5 in Semantic and 74.6 in Hand, but 36.5 in Temporal; Qwen3-VL-32B similarly scores 33.3, 75.9, and 17.1. Even the highest-scoring EgoAction model reaches only 40.1 in Temporal. Motion remains another challenge: the adapted OV2 model scores 40.1, compared with 43.4 in Verb and 47.1 in Objects. These results identify temporal extent and motion description as important targets for further improvement in dense egocentric action understanding.

5.3 Ablation Studies

Candidate-overlap threshold. To examine the sensitivity of evaluation to temporal candidate eligibility, the threshold analysis considers $\tau_{\text{sem}} \in \{0.3, 0.5, 0.7\}$, with 0.5 as the default. Predictions and pairwise semantic judgments are held fixed and the correspondences are recomputed for each threshold. The preserved ordering shows that the relative model comparisons are stable over the tested range, although absolute scores depend on the candidate-overlap requirement.

Table 4: **Overall ablation results on EgoActionScore.** Scores range from 0 to 100; higher is better. Video input sweeps use $\tau_{\text{sem}} = 0.5$. GPT-6 Astra uses fixed-count sampling.

Model	(a) Varying τ_{sem}			(b) Varying video input settings					
				Fixed-rate sampling			Fixed-count sampling		
	0.3	0.5	0.7	2 fps	5 fps	10 fps	16 frames	32 frames	64 frames
Qwen3-VL-4B	39.5	39.3	36.7	39.3	42.3	42.9	36.0	39.9	41.6
Qwen3-VL-32B	42.5	42.1	40.1	42.1	43.7	43.6	41.3	42.0	42.9
LLaVA-OV2-8B	37.5	36.7	35.1	36.7	38.8	39.1	34.8	35.8	37.5
GPT-6 Astra	62.4	61.3	56.6	—	—	—	54.9	60.2	63.0
EgoAction (Qwen3-VL-4B)	49.9	49.7	46.1	49.7	57.4	58.3	45.2	50.2	54.8
EgoAction (LLaVA-OV2-8B)	57.1	56.7	51.6	56.7	59.8	60.1	52.5	55.5	58.3

Video input settings. Dense hand-action captioning requires resolving brief actions and the transitions between them, which may be missed under sparse temporal sampling. We therefore examine whether denser visual observations enable models to describe and localize these actions more accurately over the same video interval. We consider two input settings: (a) fixed-rate sampling at 2, 5, and 10 fps, and (b) fixed-count sampling with 16, 32, and 64 uniformly sampled frames per clip. The former maintains a specified sampling rate across clips, while the latter assigns each clip a fixed frame budget. The results suggest that task-aligned adaptation complements denser visual sampling, rather than its benefits being replaced by a larger input frame budget.

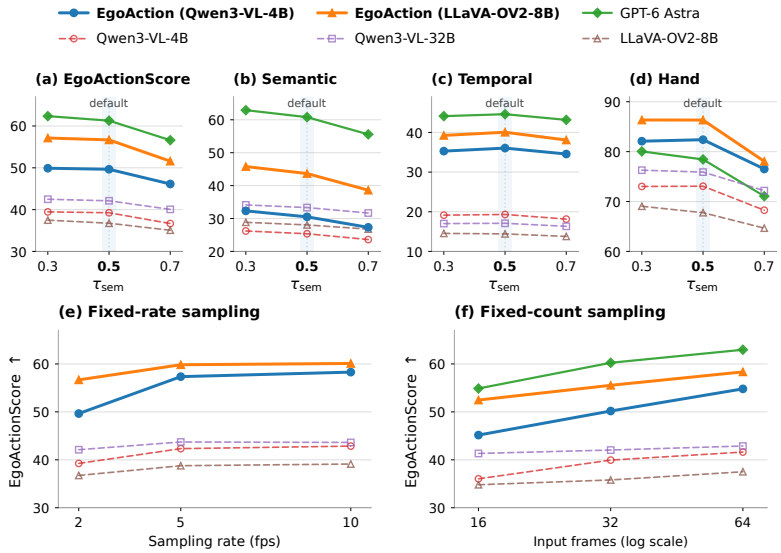

Figure 5: **Ablation analysis.** (a–d) EgoActionScore, Semantic, Temporal, and Hand under varying candidate-overlap thresholds; the shaded band marks the default of 0.5. (e–f) EgoActionScore under fixed-rate and fixed-count video sampling. Frame counts use a logarithmic axis. All scores range from 0 to 100; vertical ranges vary by metric.

6 Conclusion

We introduced EgoActionBench for evaluating dense action captioning in egocentric video. The benchmark represents manipulation through temporally localized, hand-specific action descriptions, and its evaluation suite assesses action content, temporal extent, and hand assignment on a shared correspondence. Experiments reveal a gap between general video understanding and dense action recovery. Supervised adaptation improves performance on both Qwen3-VL and LLaVA-OneVision-2, while temporal precision and motion description remain challenging. These findings establish EgoActionBench as a testbed for developing models that recover the detailed action structure of human manipulation.

AI USE STATEMENT

We used GPT-6 Astra to assist with benchmark annotation. Human reviewers verified the benchmark annotations against the original videos, whereas the labels used for supervised adaptation were generated and refined automatically. Qwen3-32B served as the text-only semantic judge described in our evaluation protocol. AI tools also assisted with code development and language polishing. AI tools were not used to originate the core research ideas or formulate the central contributions of this work. We take full responsibility for the final paper, dataset annotations, code, and reported results.

REPRODUCIBILITY STATEMENT

Section A describes the adaptation data, training settings, and semantic judge. Section B details the annotation and quality-control procedures, and section C specifies candidate retrieval, semantic judging, prediction selection, and score aggregation. The transfer evaluation setup and category-specific metrics are provided in section D.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards Next-Generation Perceptual Intelligence. *arXiv preprint arXiv:2605.25979*, 2026. URL https://arxiv.org/abs/2605.25979.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

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

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.html.

Yukang Cao, Haozhe Xie, Beichen Wen, Runmao Yao, Yinghao Liu, Yue Huang, Zhichao Liao, Yunxiang Wang, Haiheng Liu, Xingshun Tian, Dawei Su, Long Zhuo, Dacheng Tao, Xiaogang Wang, Liang Pan, and Ziwei Liu. ACE-Data-0: Human-centric ambient capture as embodied data engine, 2026. URL https://arxiv.org/abs/2607.28625.

Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. VidCapBench: A comprehensive benchmark of video captioning for controllable text-to-video generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8543–8563. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.449. URL https://aclanthology.org/2025.findings-acl.449/.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali

Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. *arXiv preprint arXiv:2601.10611*, 2026. URL https://arxiv.org/abs/2601.10611.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2. URL https://doi.org/10.1007/s11263-021-01531-2.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, Minghao Zhu, Xiao Lin, Yang Bai, Qian Jiang, Yaxi Zhao, Minghua Zeng, Junlong Gao, Yuming Jiang, Jun Cen, Siteng Huang, Liuyi Wang, Wenqiao Zhang, Chengju Liu, Jianfei Yang, Shijian Lu, and Deli Zhao. RynnBrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026. URL https://arxiv.org/abs/2602.14979.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: VIdeo segmentations and object relations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13745–13758, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/590a7ebe0da1f262c80d0188f5c4c222-Abstract-Datasets_and_Benchmarks.html.

Yufan Deng and Daquan Zhou. HumanNet: Scaling Human-centric Video Learning to One Million Hours, 2026. URL https://arxiv.org/abs/2605.06747.

Chengming Feng, Hesam Araghi, Liming Zheng, Julien Dupeyroux, Xucong Zhang, Jan van Gemert, and Nergis Tömen. Cooking beyond Frames: A Stereo Event Camera Dataset in the Kitchen. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2608.04865.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/6406_ECCV_2020_paper.php.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. DreamDojo: A Generalist Robot World Model from Large-Scale Human Videos, 2026. URL https://arxiv.org/abs/2602.06949.

Google DeepMind. Gemini 3.1 Pro model card. Official model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed 2026-09-22.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1706.04261.

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

18995–19012, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Grauman_Ego-Exo4D_Understanding_Skilled_Human_Activity_from_First-_and_Third-Person_Perspectives_CVPR_2024_paper.html.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FFxkFMU89E.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video ReCap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2402.13250.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation*, pp. 13226–13233, 2025. doi: 10.1109/ICRA55743.2025.11127989. URL https://doi.org/10.1109/ICRA55743.2025.11127989.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017. doi: 10.1109/ICCV.2017.83. URL https://doi.org/10.1109/ICCV.2017.83.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kwon_H2O_Two_Hands_Manipulating_Objects_for_First_Person_Interaction_Recognition_ICCV_2021_paper.html.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'Ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. EgoCross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):6592–6600, 2026a. doi: 10.1609/aaai.v40i8.37589. URL https://ojs.aaai.org/index.php/AAAI/article/view/37589.

Yihang Li, Xuelong Wei, Jingzhou Luo, Yingjing Xiao, Yibo Bai, Guangyuan Zhou, Teng Zou, Chenguang Gui, Jiajun Wen, He Zhang, Kangliang Chen, Xing Pan, Shuaiyan Liu, Daming Wang, Tao An, Jiayi Li, Shibo Jin, Wanwan Zhang, Tianyu Wang, Boren Wei, Zhixuan Huang, Fangsheng Liu, Ruodai Li, Hui Zhang, Anson Li, Yicheng Gong, Peng Cao, Jiaming Liang, and Liang Lin. EgoLive: A Large-Scale Egocentric Dataset from Real-World Human Tasks, 2026b. URL https://arxiv.org/abs/2604.23570.

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, Zitong Shan, Zhenchao Jin, Jiadong Hong, Taowen Wang, Yushi Feng, You Liu, Yibo Wang, Yifan Yang, Zhaowen Zhou, Man Luo, Hao Cheng, Bo Zhang, Jianshu Li, Jiansheng Cai, Guocai Yao, Jize Zhang, Chenhao Lin, Renjing Xu, Lequan Yu, Chao Shen, Chunhua Shen, and Zhe Li. Open-AoE: An Open Egocentric Manipulation Dataset and Toolchain for Embodied Learning, 2026c. URL https://arxiv.org/abs/2607.14183.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://openreview.net/forum?id=nE8_DvxAqAB.

Hui Liu and Xiaojun Wan. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*,

pp. 11807–11823, 2023. doi: 10.18653/v1/2023.emnlp-main.723. URL https://doi.org/10.18653/v1/2023.emnlp-main.723.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_TACO_Benchmarking_Generalizable_Bimanual_Tool-ACtion-Object_Understanding_CVPR_2024_paper.html.

Zhipeng Liu, Haochen Wang, and Zhaoxiang Zhang. CAPEval: A Decoupled Caption Evaluation across Understanding and Generation. *arXiv preprint arXiv:2608.02589*, 2026. URL https://arxiv.org/abs/2608.02589.

Macrodata Labs. WGO-Bench: What's going on benchmark. Hugging Face dataset card, 2026a. URL https://huggingface.co/datasets/macrodata/WGO-Bench. Accessed 2026-09-05.

Macrodata Labs. Segmenting robot video into actionable subtasks. Macrodata Labs project report and `Refiner` pipeline, 2026b. URL https://macrodata.co/blog/annotating-robot-video-subtasks. Accessed 2026-09-05.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 892–909. PMLR, 2023. URL https://proceedings.mlr.press/v205/nair23a.html.

Katsuyuki Nakamura, Hiroki Ohashi, and Mitsuhiro Okada. Sensor-augmented egocentric-video captioning with dynamic modal attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. doi: 10.1145/3474085.3475557. URL https://arxiv.org/abs/2109.02955.

Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2EgoDVC: Dense video captioning of egocentric procedural activities using web instructional videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8324–8335, 2025. doi: 10.1109/WACV61041.2025.00807. URL https://doi.org/10.1109/WACV61041.2025.00807.

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, 2026a. URL https://deploymentsafety.openai.com/gpt-5-6. Accessed 2026-09-22.

OpenAI. GPT-6 Astra system card. OpenAI Deployment Safety Hub, 2026b. URL https://deploymentsafety.openai.com/gpt-6-astra. Accessed 2026-09-22.

OpenBMB. MiniCPM-V 4.5. Official model documentation, 2025. URL https://github.com/OpenBMB/MiniCPM-V/blob/main/docs/minicpm_v4dot5_en.md.

Gunjan Paul, Senthil Palanisamy, Satpal Singh Rathore, Pratyush Kumar Patnaik, Shubhanshu Khatana, and Abhishek Anand. Ego-OSCAR: Egocentric Open source Stereo CAptuRe System, 2026. URL https://arxiv.org/abs/2608.08285.

Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, and Limin Wang. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2503.00986.

Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Yu Qiao, and Jiangmiao Pang. EgoThinker: Unveiling egocentric reasoning with spatio-temporal CoT. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2510.23569.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23901–23913, 2025. doi: 10.1109/CVPR52734.2025.02226. URL https://doi.org/10.1109/CVPR52734.2025.02226.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023. URL https://arxiv.org/abs/2307.05463.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh

Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Wenkai Xuan, Chenyu Yang, Xi Wang, James Fort, Richard Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. EgoVerse: An Egocentric Human Dataset for Robot Learning from Around the World. *arXiv preprint arXiv:2604.07607*, 2026. URL https://arxiv.org/abs/2604.07607.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.02051.

Zhaofeng Shi, Heqian Qiu, Lanxiao Wang, Qingbo Wu, Fanman Meng, and Hongliang Li. Unsupervised ego- and exo-centric dense procedural activity captioning via gaze consensus adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. URL https://arxiv.org/abs/2504.04840.

Yuejiao Su, Yi Wang, and Lap-Pui Chau. CaRe-Ego: Contact-aware relationship modeling for egocentric interactive hand-object segmentation. *Expert Systems with Applications*, 296:129148, 2025. doi: 10.1016/j.eswa.2025.129148. URL https://doi.org/10.1016/j.eswa.2025.129148.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. URL https://arxiv.org/abs/2506.15220.

Masatoshi Tateno, Gido Kato, Hirokatsu Kataoka, Yoichi Sato, and Takuma Yagi. HanDyVQA: A video QA benchmark for fine-grained hand-object interaction dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3455–3465, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tateno_HanDyVQA_A_Video_QA_Benchmark_for_Fine-Grained_Hand-Object_Interaction_Dynamics_CVPR_2026_paper.html.

Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-Only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5250–5261, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Ego-Only_Egocentric_Action_Detection_without_Exocentric_Transferring_ICCV_2023_paper.html.

Jiahao Wang, An Ping, Yanghai Wang, Yuanxing Zhang, Shihao Li, Hanyan Bian, Yichi Ren, Yize Zhang, Han Wang, Haowen Chen, Junze Li, Jiaqi Wang, Yiyang Hu, Zhuze Xu, Zijie Zhang, and Jiaheng Liu. OmniCap-IF: Benchmarking and improving instruction following abilities for omni-video captioning. *arXiv preprint arXiv:2606.08572*, 2026a. URL https://arxiv.org/abs/2606.08572.

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, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Ye Wang, Pei Lin, Xiong-Hui Chen, Haoqi Yuan, Zhixuan Liang, Yiyang Huang, Anzhe Chen, Zixing Lei, Jie Zhang, Tao Zhang, Haoyang Li, Tong Zhang, Chenxi Xiao, Ziyuan Jiao, and Qin Jin. Ego2Robot: Scalable robot data synthesis from egocentric human data, 2026b. URL https://arxiv.org/abs/2608.02580.

Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. HumanEgo: Zero-Shot Robot Learning from Minutes of Human Egocentric Videos, 2026c. URL https://arxiv.org/abs/2605.24934.

Boshen Xu, Yuting Mei, Xinbi Liu, Sipeng Zheng, and Qin Jin. EgoDTM: Towards 3d-aware egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*,

volume 38, 2025. doi: 10.52202/085713-4088. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b1a81aa84530d9fa99e420b2305cc11a-Abstract-Conference.html.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016. doi: 10.1109/CVPR.2016.571. URL https://doi.org/10.1109/CVPR.2016.571.

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

Penghui Yang, Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Yibin Wang, Yujie Zhou, Jiazi Bu, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. CapRL++: Unified reinforcement learning with verifiable rewards for dense image and video captioning. *arXiv preprint arXiv:2606.09393*, 2026. URL https://arxiv.org/abs/2606.09393.

Linli Yao, Yuancheng Wei, Yaojie Zhang, Lei Li, Xinlong Chen, Feifan Song, Ziyue Wang, Kun Ouyang, Yuanxin Liu, Lingpeng Kong, Qi Liu, Pengfei Wan, Kun Gai, Yuanxing Zhang, and Xu Sun. TimeChat-Captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.08711.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MM-Ego: Towards building egocentric multimodal LLMs for video QA. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07177.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025. URL https://arxiv.org/abs/2501.13106.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 2022a. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/7278_ECCV_2022_paper.php.

Haoyu Zhang, Qiaohui Chu, Meng Liu, Haoxiang Shi, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided MLLM for egocentric video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12502–12510, 2026. doi: 10.1609/aaai.v40i15.38244. URL https://ojs.aaai.org/index.php/AAAI/article/view/38244.

Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pp. 127–145, 2022b. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136890125.pdf.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2212.04501.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling dexterous manipulation with diverse egocentric human data, 2026. URL https://arxiv.org/abs/2602.16710.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018a. doi: 10.1609/aaai.v32i1.12342. URL https://doi.org/10.1609/aaai.v32i1.12342.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8739–8748, 2018b. URL https://arxiv.org/abs/1804.00819.

A Additional Experimental Details

Adaptation data. The 5,354 automatically annotated episodes comprise 5,146 training examples and 208 validation examples. Source-video identifiers separate this corpus from EgoActionBench. The benchmark annotations undergo human verification; the adaptation labels are generated and refined automatically.

Supervised adaptation. We use a learning rate of 10^{-5}, a cosine schedule, and an effective batch size of 32. Training runs for three epochs; the evaluated models are selected after the second epoch using validation loss. As described in section 4, the visual encoders remain frozen, while the language models and multimodal connectors are updated. Only the target response contributes to the next-token prediction loss.

Semantic evaluation. Qwen3-32B serves as a text-only judge with temperature zero. It receives the reference caption, its structured semantic fields, and the predicted caption, without video frames. Every temporally eligible candidate is evaluated. The default candidate-overlap threshold is $\tau_{\text{sem}} = 0.5$; selection and score aggregation follow section C.

B Data Annotation and Quality Control

This appendix details the five-stage annotation pipeline introduced in section 3.3. Model-generated captions serve as initial drafts, while the original video provides the evidence for human annotation, agentic quality control, and final human verification.

Stage 1: Hand-aware preprocessing. We extract left and right hand masks using CaRe-Ego (Su et al., 2025). Temporal smoothing filters brief or inconsistent detections using the continuity of hand regions across neighboring frames. The filtered masks are rendered as distinct overlays that highlight each hand while preserving the surrounding objects and scene. This hand-highlighted video provides explicit visual cues for the subsequent per-hand captioning.

Stage 2: Sliding-window captioning. A VLM processes the hand-highlighted video in sliding windows and produces one-second caption proposals for the Left, Right, and Both tracks. Frames within the current window provide visual evidence, while recent captions supply context for continuity. Because a continuous action may span multiple windows, proposals are merged separately within each track. Temporal adjacency and text similarity identify potential duplicates; the corresponding video evidence is then checked to determine whether these fragments describe the same uninterrupted operation. Task-relevant changes in the operation, manipulated object, or contact relation remain separate. The resulting variable-length spans form the drafts for human annotation.

Stage 3: Human annotation. Annotators inspect the original video and revise the drafts according to the task definition in section 3.1. They correct action descriptions and hand assignments, add omitted actions, split compound events, and remove unsupported claims. Adjacent fragments are merged when they describe the same uninterrupted action. The three tracks are reviewed independently, allowing one hand to maintain an action while the other performs several shorter operations. Native decoded frames around transitions are inspected to refine action start and end times.

Stage 4: Agentic quality control. Reviewing agents identify intervals where descriptions or boundaries require closer examination and inspect timestamped frames extracted from those intervals. When a transition remains ambiguous, neighboring native frames are examined at finer temporal granularity. Three complementary visual cues guide this review:

- **Hand-object contact:** changes in contact or control.
- **Hand-motion phase:** changes between successive operations, including those performed under a stable grasp.
- **Object-state change:** visible responses such as displacement, deformation, or content transfer.

These cues support checks of both caption content and temporal boundaries. Caption revisions address errors in the operation, participating objects, or hand roles. Boundary revisions target the first visible onset of a task-relevant transition, rather than a later frame where its outcome is already apparent. The cues need not change simultaneously: one may support a boundary while another remains uninformative. Each proposed revision retains its supporting frame references and review rationale; unresolved cases are flagged for human verification.

Stage 5: Final human verification. Human reviewers check the annotations against the original video, using the proposed revisions, supporting frames, and rationales from Stage 4. They resolve remaining ambiguities and determine whether further correction is needed. Episodes requiring rework return to Stage 3 and undergo agentic QC again. Annotations are finalized only after human verification.

C Evaluation Protocol

This appendix specifies the evaluation suite introduced in section 3.4. Matching is performed independently for each ground-truth action. A single selected prediction supplies all three component scores, preventing different predictions from contributing separately to the content, temporal, and hand assessments of the same action.

Evaluation inputs. Each ground-truth action and prediction are represented as

$$g = (I_g, h_g, c_g, F_g), \qquad p = (I_p, \hat{h}_p, \hat{c}_p), \tag{2}$$

where I is the temporal interval, $h \in \{L, R, B\}$ is the acting hand, and c is the natural-language caption. The structured description F_g is provided only on the ground-truth side for semantic evaluation. Models submit an interval, a hand track, and a caption; they are not required to predict structured fields.

Stage 1: Candidate retrieval. For each ground-truth action g, candidates are drawn from predictions in the same video. The admissible hand tracks are

$$\mathcal{H}(h_g) = \begin{cases} \{L, R\}, & h_g \in \{L, R\}, \\ \{B\}, & h_g = B. \end{cases} \tag{3}$$

A single-hand action can retrieve predictions from either single-hand track, allowing an otherwise correct prediction to remain eligible despite a left–right attribution error. Bimanual and single-hand actions are kept separate.

Temporal eligibility is determined by overlap normalized by the shorter interval:

$$O_{\min}(g, p) = \frac{|I_g \cap I_p|}{\min(|I_g|, |I_p|)}. \tag{4}$$

Let $\mathcal{P}$ denote the predictions in the current video. The eligible candidate set is

$$C(g) = \left\{ p \in \mathcal{P} \;\middle|\; \hat{h}_p \in \mathcal{H}(h_g),\ O_{\min}(g, p) \geq \tau_{\text{sem}} \right\}. \tag{5}$$

The default setting uses $\tau_{\text{sem}} = 0.5$ and evaluates every eligible candidate; threshold sensitivity is considered in section 5.3. This criterion admits both a long prediction containing a short ground-truth interval and a short prediction contained within a long ground-truth interval. It is used only for retrieval: predictions extending beyond or covering only part of the ground-truth interval receive lower temporal IoU scores during component scoring.

Stage 2: Semantic judging. The structured reference $F_g = (v_g, O_g, U_g, \mu_g)$ contains four fields: **Verb**, the core operation; **Objects**, the participating entities and their action-dependent roles; **Tools**, the intermediaries used to perform the operation; and **Motion**, the coarse physical movement underlying the operation. Objects and Tools are distinguished by their role: a knife is a tool when used for cutting, but an object when it is picked up. Required Motion values belong to `reach`, `translate`, `lift`, `lower`, `rotate`, `press`, `oscillate`, or `static`.

For every $p \in C(g)$, a language-model judge receives the ground-truth caption c_g, structured fields F_g, and full predicted caption $\hat{c}_p$, with no access to video frames. Required content is assessed using direct quotes from the prediction. Equivalent wording receives credit, while missing or contradictory content does not. Supporting evidence must refer to the same action and cannot be assembled across distinct actions in the caption.

Verb and Motion receive binary scores. Objects and Tools are evaluated item by item and averaged within each field. The ground-truth annotation fixes which fields contribute: fields with required content or explicitly annotated absence are included, while unscored and inapplicable fields are excluded. Required content omitted from a prediction receives zero credit and remains in the denominator.

Let K_g denote the scored fields and $r_k(g,p) \in [0,1]$ their individual scores. All reported runs use equal field weights:

$$S(g,p) = \frac{1}{|K_g|} \sum_{k \in K_g} r_k(g,p). \tag{6}$$

A field therefore receives no additional weight for containing more items. For every eligible pair, the judge must resolve each required field and its action binding.

Selecting one prediction per ground-truth action. For each nonempty candidate set, we select the prediction with the highest semantic score:

$$p^*(g) = \arg\max_{p \in C(g)} S(g,p). \tag{7}$$

Exact ties are broken by higher temporal IoU, agreement with the ground-truth hand, earlier start, earlier end, and prediction ID, in that order. Each ground-truth action selects at most one prediction, but the same prediction may be selected for multiple ground-truth actions. We retain the selected caption and interval without combining evidence across predictions or cropping temporal boundaries.

Stage 3: Component scoring. Using the selected prediction, we compute semantic agreement, temporal IoU, and hand-label agreement:

$$(S_g, T_g, H_g) = \begin{cases} \left(S(g,p^*(g)), \dfrac{|I_g \cap I_{p^*(g)}|}{|I_g \cup I_{p^*(g)}|}, \mathbf{1}[h_g = \hat{h}_{p^*(g)}] \right), & C(g) \neq \varnothing, \\ (0,0,0), & C(g) = \varnothing. \end{cases} \tag{8}$$

All three scores use the same prediction. Temporal and Hand are computed independently of whether its Semantic score is positive. A ground-truth action with no eligible candidate receives three zeros and remains included in dataset-level averaging.

Dataset-level aggregation. Let $\mathcal{G}$ denote all ground-truth actions. Each action receives equal weight, and component scores are reported on a 0–100 scale:

$$(\text{Semantic}, \text{Temporal}, \text{Hand}) = \frac{100}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} (S_g, T_g, H_g). \tag{9}$$

Their arithmetic mean defines the aggregate score:

$$\text{EgoActionScore} = \frac{\text{Semantic} + \text{Temporal} + \text{Hand}}{3}. \tag{10}$$

We additionally report scores for each of the four semantic fields.

Interpretation. EgoActionScore measures recovery of ground-truth actions. It does not measure prediction precision or F1: extra predictions incur no independent false-positive penalty, and one prediction may support multiple ground-truth actions. Hand scores should also be interpreted in light of candidate filtering. Since a Both ground-truth action admits only Both predictions, its Hand score reflects whether an eligible candidate is available, rather than unrestricted three-way hand classification.

D TRANSFER TO HANDYVQA

To examine whether dense action supervision transfers beyond caption generation, we evaluate Qwen3-VL-4B and its EgoAction adaptation directly on HanDyVQA (Tateno et al., 2026), without additional VQA fine-tuning. The evaluation comprises 8,348 questions from recordings disjoint from our adaptation training and validation data. Both models receive identical video inputs and follow the official question prompts and category-specific metrics. Accuracy is reported for Action, Process, Location, State, and Parts, and average precision (AP) for Objects.

Table 5: **Transfer to HanDyVQA.** Qwen3-VL-4B and its EgoAction adaptation are evaluated without additional VQA fine-tuning on 8,348 questions from recordings disjoint from the adaptation data. Accuracy is reported for the five single-answer categories and average precision for Objects. Scores are on a $[0, 100]$ scale. Bold indicates the better result within each category.

Category	Metric	Questions	Qwen3-VL-4B	EgoAction (Qwen3-VL-4B)	Δ
Action	Acc	1,450	70.07	**71.93**	+1.86
Process	Acc	1,372	60.06	**64.14**	+4.08
Location	Acc	1,425	52.77	**56.84**	+4.07
State	Acc	1,411	62.93	**64.14**	+1.20
Parts	Acc	1,362	58.96	**61.75**	+2.79
Objects	AP	1,328	**59.54**	55.68	-3.85
Single-answer pooled	Acc	7,020	61.01	**63.80**	+2.79

Table 5 shows improvements in all five single-answer categories. Pooled accuracy increases from 61.01% to 63.80%, with the largest gains in Process, from 60.06% to 64.14%, and Location, from 52.77% to 56.84%. Action, State, and Parts also improve. Objects AP decreases from 59.54 to 55.68. Dense action supervision therefore transfers positively to several hand-object reasoning categories, while the Objects category exhibits a different response to adaptation.

E INSIGHTS AND DISCUSSION

Strong VLMs still struggle with precise action recovery. GPT-6 Astra achieves the highest model EgoActionScore of 61.3, yet remains 31.2 points below the human baseline. The largest component gap is in **temporal localization**: GPT-6 scores 44.6 compared with 92.2 for humans, indicating that accurately delimiting individual actions remains a major challenge. Semantic recovery is also incomplete. Although GPT-6 reaches 80.4 on *Tools*, its *Verb* and *Motion* scores are only 63.3 and 48.9, respectively. These results suggest that identifying the instruments involved in an action is more reliable than recovering the operation and its underlying movement. Strong overall performance therefore leaves substantial room for improvement in both action content and temporal precision.

Supervision at the action level. Supervised fine-tuning improves all three evaluation components on both backbones. For LLaVA-OV2-8B, Semantic increases from 28.1 to 43.7, Temporal from 14.4 to 40.1, and Hand from 67.8 to 86.3. For Qwen3-VL-4B, the corresponding scores increase from 25.4 to 30.5, 19.3 to 36.1, and 73.1 to 82.4. Temporal localization improves most on both backbones, suggesting that action-level supervision is particularly useful for learning when an operation occurs. Improvements in hand assignment further support learning these aspects together. Even after adaptation, Temporal and Motion scores reach only about 40 at best, leaving substantial room for more precise action recovery.

Action content and temporal extent. Figure 6 illustrates why caption content and temporal extent should be examined together. In the carpenter-square case, a long prediction describes actions that correspond to multiple shorter references. Its content can support those references, while its full interval exposes the loss of temporal precision. The tomato case illustrates the importance of retaining the caption and interval of a single selected prediction, rather than constructing an answer by merging different predictions. Together, these examples distinguish recovering an action's meaning from recovering its temporal granularity, complementing the component-wise analysis in Section 5.2.

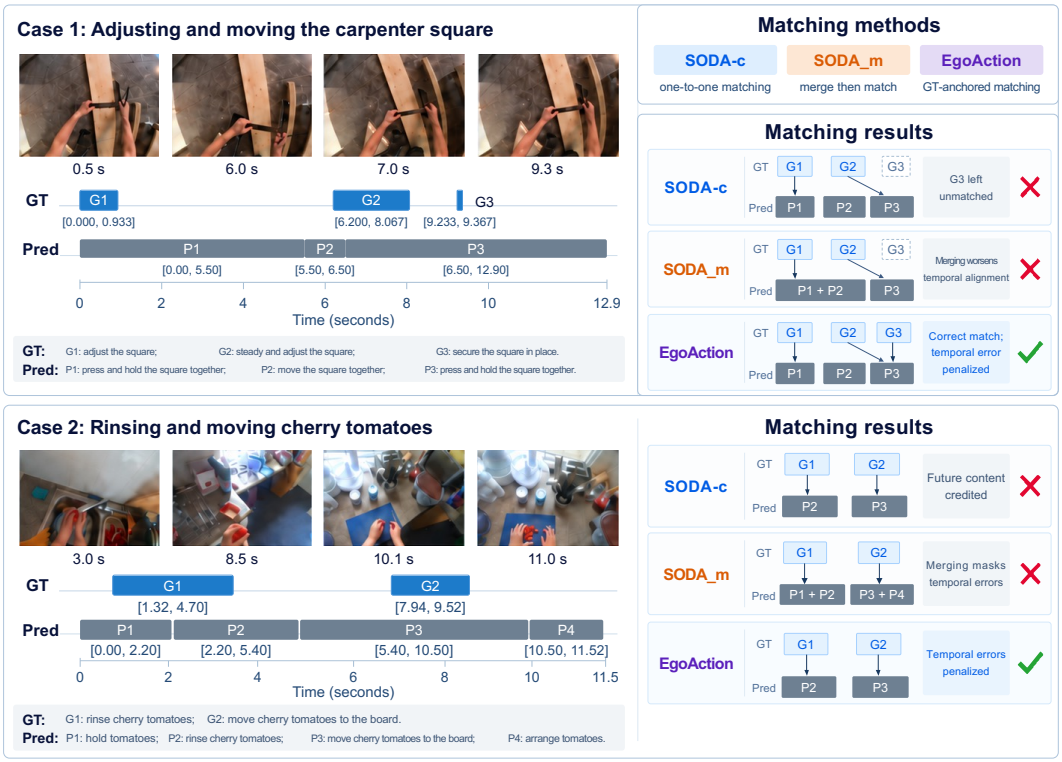

Figure 6: **Qualitative Comparison of Action-Matching Strategies.** Two illustrative cases compare one-to-one, merge-then-match, and GT-anchored correspondence. The carpenter-square example highlights reuse of a long prediction for multiple short reference actions; its full interval remains the basis for temporal scoring. The tomato example illustrates selecting a single prediction for each reference and keeping its caption and interval intact. Intervals and caption text are shown alongside the example correspondences.