EGOGAPBENCH: BENCHMARKING EGOCENTRIC ACTION SELECTION IN MULTI-AGENT SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) deployed in devices such as smart glasses need to identify appropriate next actions from the user's perspective. However, some questions in existing evaluations based on first-person visual inputs can be answered based solely on the camera wearer's body cues. When other people's actions are absent from the answer choices, a correct response provides limited evidence that the model distinguishes the user's situation from those of others, so it is difficult to assess action selection from the user's perspective separately from input understanding. To address this limitation, we use images of multiple people that contain no body cues from the camera wearer and ask models to respond as if they were directly viewing the scene. We define Egocentric Action Selection (EAS) as the task of selecting an appropriate action from the user's perspective in a given situation and introduce EgoGapBench, a diagnostic benchmark for EAS with 1,000 multiple-choice questions on 440 COCO images. EgoGapBench includes other people's actions as distractors when duplicating those actions would be inappropriate in the given situation. All evaluated open- and closed-source MLLMs achieve substantially lower accuracy than the human baseline, and for most models errors concentrate on the actions of other people in the scene. Moreover, after instruction tuning on first-person visual input data, all five open-source models score nearly the same on an existing benchmark of first-person visual inputs (-0.5% relative) but drop on EgoGapBench (-33.9% relative).

1 INTRODUCTION

Multimodal large language models (MLLMs) (Alayrac et al., 2022; Liu et al., 2023; Dai et al., 2023) are increasingly used through AI glasses to share the scene the user sees and to assist the user (Jiang et al., 2026). Current products answer questions about what is in view, give directions, or show step-by-step instructions (Meta, 2025; Google, 2026). Because the action is performed by the user, the model has to select an action that fits the user's situation. For example, suppose that a passenger wearing AI glasses is watching the driver hold the steering wheel (Figure 1). Turning the wheel is an appropriate action for the driver, but if the model suggests it to the passenger and the passenger follows the suggestion, it can interfere with the driver.



Figure 1: Illustrative scenario of third-to-first-person confusion: the model suggests the driver's action to the user.

Such third-to-first-person confusion, between the third-person perspective on other people and the user's first-person perspective, leads the model to recommend actions that are inappropriate for the user. Even when the model shares the user's view, other people in the scene remain objects of observation. Yet the model's response can suggest to the user what those people are doing, or what they would do, confusing their perspective with the user's.

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.

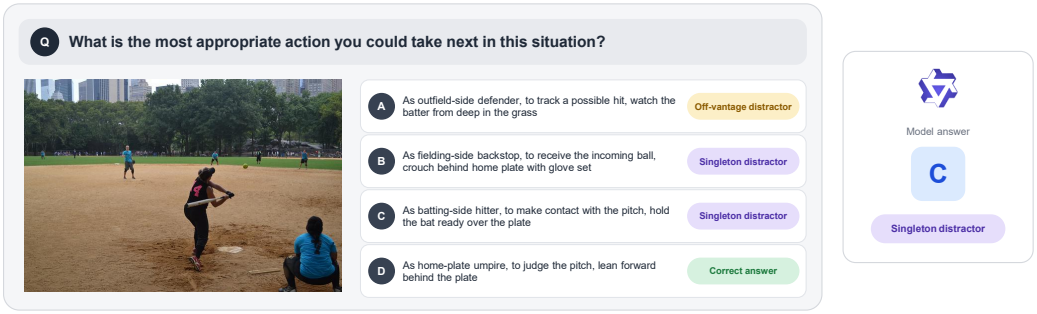

Figure 2: An example question from EgoGapBench. The viewer is behind home plate. Besides the correct answer (D), the choices include two *singleton distractors*, actions of other people in the scene (B, C), and one *off-vantage distractor*, an action that cannot be performed from the viewer's position (A). Qwen3.5 selects C, the batter's action.

Existing benchmarks evaluate understanding of the first-person perspective (Cheng et al., 2024; Grauman et al., 2022) and the third-person perspective (Hu et al., 2024) separately, and some cover both (Grauman et al., 2024). However, whether models confuse the two perspectives has rarely been measured. Measuring this confusion with existing egocentric action-selection benchmarks poses two difficulties. First, the camera wearer's body is mostly visible in these benchmarks (Grauman et al., 2022; Damen et al., 2020). Since models may rely on hand cues (Tateno et al., 2026) and some questions may be answerable from hand-object interaction alone (Chen et al., 2026), in such cases it is hard to tell whether the model distinguished the user's perspective from that of others. Second, existing egocentric datasets focus mainly on a single person's activities, as prior work has pointed out (Sakai et al., 2025; Yang et al., 2025), and even when other people appear, their actions are not among the answer choices (Rezaei et al., 2025). In EgoNormia, a benchmark of norm-appropriate action selection, for example, distractors are actions of the wearer that are appropriate only in altered contexts rather than actions of other people. Replacing one choice with another person's action from the same scene substantially reduced Gemini-3-Flash's (Google DeepMind, 2025) accuracy, whereas a random action did not (Appendix B).

To diagnose this confusion in action selection, we define a task in which the model selects an action appropriate to perform immediately, as if it were the user directly viewing the depicted scene. We call this task Egocentric Action Selection (EAS). In this setting, we use images that depict multiple people and contain no body cues from the camera wearer, and we include other people's actions in the scene among the answer choices so that the model must judge whether those actions are also appropriate for the user. To evaluate EAS, we introduce EgoGapBench, comprising 1,000 action-selection questions over 440 COCO images (Figure 2). Since multiple actions may be appropriate in a scene, each question pairs one appropriate action with clearly inappropriate distractors. To construct these distractors, we select images in which two people's actions each satisfy a singleton constraint: the action depends on a role that only one person can hold or an object that only one person can control, so performing it as well is inappropriate for the user. Correct actions range from direct participation to indirect responses such as watching, and some questions add a distractor that cannot be performed from the user's vantage point.

All open- and closed-source MLLMs evaluated on EgoGapBench score below the human baseline. Even GPT-5.4, the most accurate model, reaches 66.1% against a human baseline of 94.5%. The gap is largest on questions whose correct answer is direct participation, and for most models errors concentrate on the actions of other people in the scene. After fine-tuning on EgoIT-99K (Yang et al., 2025), an instruction-tuning dataset built on first-person visual inputs, all five open-source models show lower EgoGapBench accuracy, with a mean relative change of -33.9%, while the mean relative change on EgoThink (Cheng et al., 2024) is only -0.5%. Fine-tuning on an EAS training split built from images excluded from EgoGapBench does raise accuracy, but the gap to the human baseline remains, and some errors shift from the actions of other people to actions that cannot be performed from the user's vantage point.

Contributions. Our contributions are as follows:

- We define EAS, a task that diagnoses third-to-first-person confusion in action selection.
- We construct EgoGapBench, comprising 1,000 questions over 440 images, using the singleton constraint and human verification.
- We show that evaluated MLLMs perform substantially below the human baseline and frequently select actions of other people that are inappropriate for the user.

2 Related Work

2.1 Egocentric Scene Understanding

Existing egocentric benchmarks evaluate understanding of first-person visual inputs. Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2020) provide large-scale video datasets for recognizing activities, objects, and interactions, as well as predicting future actions. Based on these data, EgoThink (Cheng et al., 2024) evaluates perception and reasoning on first-person visual inputs, while EgoSchema (Mangalam et al., 2023) evaluates long-video understanding. These benchmarks assess models' understanding of and reasoning about the camera wearer's activities and surroundings. EgoGapBench evaluates action selection from the user's egocentric perspective separately from the input format.

2.2 Action Selection in MLLMs

Several benchmarks evaluate models' ability to plan or select the next action from first-person or third-person visual inputs. EgoPlan-Bench (Chen et al., 2026) evaluates action planning from video, using the action that actually follows as the correct answer and actions from other steps of the same task as distractors. EgoNormia (Rezaei et al., 2025) uses a norm-appropriate action in the observed context as the correct answer and actions that are norm-appropriate in modified contexts but not the most appropriate in the observed context as distractors. VIVA (Hu et al., 2024) evaluates action selection for a depicted situation on third-person web images, with distractors that worsen the situation or hold only under specific conditions. EgoGapBench uses an action appropriate for the user in the given situation as the correct answer and other people's observed actions that are inappropriate for the user as distractors.

2.3 Reasoning from User and Agent Perspectives

Prior work evaluates models' ability to interpret scenes and make decisions from a user's or agent's perspective. EgoOrientBench (Jung et al., 2025) evaluates MLLMs' understanding of object orientation from the user's egocentric perspective on images from datasets such as ImageNet (Deng et al., 2009) and PACS (Li et al., 2017), using the user as the reference for interpretation. MindPower (Zhang et al., 2026) assigns the model a robot role through textual instructions and asks it to infer a person's state and select actions to assist that person. EgoGapBench instead asks the model to select an appropriate action for a human user observing other people. Because the user and others can perform the same kinds of actions, the model must distinguish the user's perspective from those of other people to judge whether an action is appropriate for the user.

3 EgoGapBench

3.1 Task Definition

We define EAS to diagnose third-to-first-person confusion in action selection. In EAS, the model is given an image of a scene with multiple people, a question, and a set of answer choices, and is asked to select an action as if it were the user directly viewing the depicted scene from the camera's position. We refer to this user as the viewer. Each answer choice is a sentence that combines a role, a goal, and an action, making explicit in which role and for what goal the action is performed. The correct answer is an action that the viewer can begin immediately from the current position, without first relocating or waiting for a turn, judged together with the role and goal stated in the choice. Each

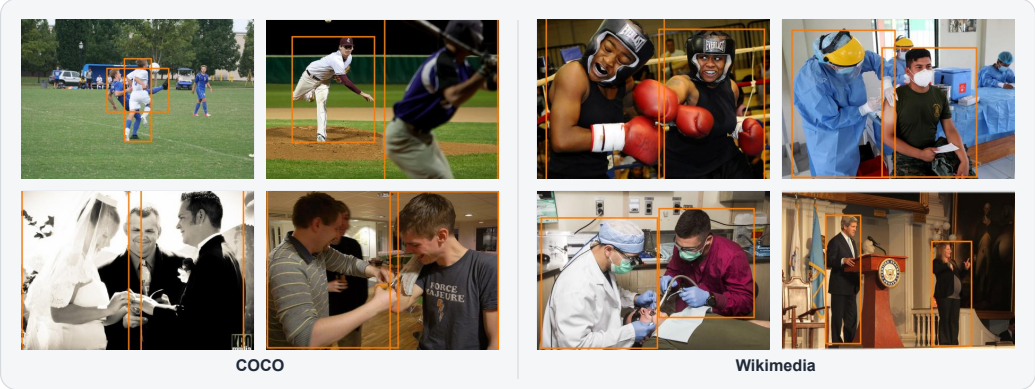

Figure 3: Example images that satisfy the singleton constraint, from COCO (left) and Wikimedia (right). Seen from the camera's position, the roles or objects in each scene are already taken by the people shown, so the viewer cannot take them as well. Bounding boxes mark the two people, each in such a role, whose actions serve as distractors.

question has exactly one correct answer. Because appropriate actions range from direct participation to indirect responses, we label questions by type to compare accuracy across them. A question is direct when its correct answer involves direct participation and indirect when the correct answer is an indirect response such as watching or listening. Each question has either three (n3) or four (n4) answer choices. An n3 question contains one correct answer and two singleton distractors: observed actions of other people in the scene that, under the singleton constraint, are inappropriate for the user to perform as well because the role or object is already taken (Section 3.2). An n4 question adds one off-vantage distractor: an action that cannot be performed immediately from the viewer's current position, regardless of whether another person is performing it. In Figure 2, for example, watching the batter from deep in the outfield (option A) cannot be done from behind home plate.

3.2 Design Principles

EgoGapBench is built on two design principles.

Principle 1: Others present, no camera wearer's body. We use third-person images in which multiple people appear and no camera wearer's body is visible. This removes the viewer's hands, feet, and other body movements as cues for action selection. In addition, we include other people's observed actions among the answer choices so that the model must determine whether those actions are also appropriate for the viewer.

Principle 2: Singleton distractors. Many actions observed from other people can also be performed appropriately by the user, so an observed action alone does not make an unambiguous distractor. We assess each choice's role, goal, and action together in the scene context, rather than requiring each component to be exclusive to one person. We use a choice as a distractor only when another person is already performing that combination and the singleton constraint makes it inappropriate for the viewer to perform it as well. For example, in a baseball scene where a batter is already at the plate, preparing to bat as a second batter is inappropriate in that game situation, so a choice containing "hold the bat ready over the plate" is a singleton distractor (Figure 2).

3.3 Dataset Construction

To implement the two principles, we build EgoGapBench in four steps. We select singleton scenes from COCO train2017 (Lin et al., 2014), generate answer choices for each scene, verify them with human annotators, and assemble the final questions. Per-stage counts and the models used are given in Appendix D.

Image selection. Image datasets such as COCO do not annotate whether the roles or actions of people in a scene satisfy the singleton constraint. We therefore use MLLMs to identify objects or roles that satisfy the singleton constraint, such as the bride and groom at a wedding or the player in possession of the ball in a soccer match (Figure 3). For COCO images containing at least two people, Qwen3.5-35B-A3B (Qwen Team, 2026) generates annotations of singleton objects or roles and the action of the person holding each. Since a person in a singleton role may also perform everyday actions that anyone could perform, we keep images in which two different people each perform an action tied to their singleton role or object, and re-check them with gpt-5.4-mini-2026-03-17 (OpenAI, 2026). To reduce concentration on particular activities, we keep at most 50 images per activity category. Two human annotators then independently check whether each image's singleton structure is valid and whether the associated action descriptions match the scene. Only images approved by both annotators are retained.

Choice generation. For each retained image, gpt-5.4-2026-03-05 (OpenAI, 2026) generates correct-answer candidates and off-vantage candidates. We request two to four correct-answer candidates per image: one to three direct and one indirect. We request more direct candidates because, in initial generation rounds, direct candidates were rejected in verification more often than indirect ones. We also request one to two off-vantage candidates per image. Singleton distractors are rewritten with the same model from the observed actions identified during image selection. To reduce selection bias due to phrasing differences, all choices are written in the form "As [role], to [goal], [action]."

Annotation. Two annotators independently label each candidate. A correct-answer candidate receives one of three labels: valid correct answer, off-vantage, or discard. The off-vantage label is used when a candidate is not appropriate for the viewer's position but qualifies as an off-vantage distractor, so that it is reused rather than discarded. An off-vantage candidate, generated specifically to supply n4 questions, receives one of two labels: valid off-vantage or discard. We retain only candidates on which both annotators agree on a valid label. Candidates labeled discard by either annotator or assigned different labels are removed. For the 1,686 correct-answer candidates, Cohen's κ is 0.66 for the binary judgment of answer validity, with off-vantage and discard grouped as one label. Appendix E reports three-way agreement and agreement for off-vantage candidates.

Question assembly. Each verified correct answer becomes one question. After verification, we split the data at the image level: questions from 440 images form the 1,000-question benchmark, and the remaining 168 questions from 73 images form the EAS training split used in the fine-tuning analysis, so that no image appears in both (Appendix D). Each correct action is combined with two singleton distractors. When a verified off-vantage candidate exists, the question is assembled in the n4 format, and otherwise in the n3 format. The order of choices in each question is randomly shuffled once at assembly and fixed for all evaluations. The benchmark contains 650 direct and 350 indirect questions, and 618 n3 and 382 n4 questions. Detailed statistics on candidate labels are given in Appendix E.

Out-of-domain set. To test whether the results hold beyond COCO, we apply a similar pipeline to 44 manually collected Wikimedia images and construct an out-of-domain set of 107 questions (Appendix F).

4 Experiments

4.1 Setup

Evaluation. Accuracy is computed by exact string match between the predicted choice and the correct answer, with unparseable responses counted as incorrect (Appendix C). We report overall accuracy and accuracy on direct and indirect questions. The chance level for overall accuracy is 30.2%, the average of 1/3 and 1/4 weighted by the numbers of n3 and n4 questions. For error analysis on n4 questions, we report two measures. The singleton share is the number of responses that select a singleton distractor divided by the number of responses that select any distractor. The selection rate of each distractor type is the number of responses that select that type divided by the number of n4 questions.

Table 1: Zero-shot results on EgoGapBench (%), sorted by overall accuracy. Sel. rate: fraction of all n4 questions on which a singleton (Sing.) or off-vantage (Off-v.) distractor is selected. Sing. share: among n4 responses that select a distractor, the fraction that select a singleton distractor (66.7% under uniform selection). gpt-5.4 and gemini-3-flash use low reasoning effort unless stated otherwise.

	Accuracy (%)					Sel. rate (% of n4)		Sing. share
Model	Overall	n3	n4	Direct	Indirect	Sing.	Off-v.	(% of n4 err.)
Chance	30.2	33.3	25.0	30.1	30.2	50.0	25.0	66.7
gpt-5.4 (medium)	66.1	66.3	65.7	50.9	94.3	21.5	12.8	62.6
gpt-5.4	58.3	58.9	57.3	40.9	90.6	32.7	9.9	76.7
gemma4	51.7	53.4	49.0	34.0	84.6	35.3	15.7	69.2
gemini-3-flash	39.2	42.7	33.5	27.8	60.3	58.9	7.6	88.6
qwen35	29.0	31.9	24.3	20.0	45.7	69.4	6.3	91.7
qwen25vl	28.6	29.6	27.0	21.8	41.1	63.4	9.7	86.7
gpt-4o	25.8	31.1	17.3	11.8	51.7	75.4	7.3	91.1
internvl35	24.1	25.1	22.5	20.5	30.9	74.6	2.9	96.3
internvl25	16.1	15.0	17.8	17.2	14.0	77.7	4.5	94.6
Human[†]	94.5	99.2	86.8	93.2	97.1	4.8	8.3	36.2

[†]Mean of per-annotator values over three annotators (94.0–95.0 overall) who did not take part in data construction, each answering the same 200 randomly sampled questions (132 direct and 68 indirect, 124 n3 and 76 n4).

Models. We evaluate open- and closed-source MLLMs in a zero-shot setting. The open-source models are gemma-4-E4B-it (Google DeepMind, 2026), Qwen2.5-VL-7B-Instruct (Qwen Team, 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL2.5-8B (Chen et al., 2024), and InternVL3.5-8B (Wang et al., 2025). The closed-source models are gpt-5.4-2026-03-05 (OpenAI, 2026), gpt-4o-2024-11-20 (Hurst et al., 2024), and gemini-3-flash-preview (Google DeepMind, 2025). gpt-5.4 is evaluated at two reasoning-effort settings, low and medium, which we count as separate configurations, giving nine configurations in total, and gemini-3-flash-preview is run at low reasoning effort. All open-source models are run without thinking mode, with greedy decoding (temperature 0) and a fixed seed (42), and each question is evaluated once. The human baseline is the mean over three annotators who did not participate in benchmark construction, each answering the same 200 randomly sampled questions.

Fine-tuning. We fine-tune the open-source models separately on two datasets. The first is EgoIT-99K (Yang et al., 2025), an instruction-tuning dataset built on first-person visual inputs. The second is the EgoGapBench training split, in which each question is paired with a short rationale generated by gpt-5.4 given the correct answer. This setting tests whether training on our annotated data improves EAS performance on held-out images. The fine-tuned model receives only the image, question, and answer choices, and is trained to output the correct answer and rationale. To check whether the same fine-tuning affects understanding of first-person visual inputs, we also evaluate the models fine-tuned on EgoIT-99K on EgoThink, an image-based benchmark like EgoGapBench. Training details, the rationale generation prompt, and the EgoThink evaluation protocol are given in Appendix G.

4.2 Models Fail at Egocentric Action Selection

Result on EgoGapBench. As shown in Table 1, all evaluated MLLMs score below the human baseline on EgoGapBench. The human baseline is 94.5%, whereas GPT-5.4 (medium), the most accurate configuration, reaches 66.1%. Most open-source models score near or below the chance level of 30.2%.

Accuracy is lower on direct questions than on indirect questions in all configurations except InternVL2.5-8B. GPT-5.4 (medium), for example, reaches 94.3% on indirect questions but 50.9%

Table 2: Fine-tuning on EgoIT-99K. EgoThink is the mean of $\{0, 0.5, 1\}$ judge grades (gpt-4o as judge; Appendix G), and EgoGapBench is accuracy (%). Since the two scales differ, rel. gives the relative change from zero-shot (ZS) to after fine-tuning (EgoIT) per model, and the Mean row averages the per-model values.

	EgoThink (0–1)			EgoGapBench (acc. %)		
Model	ZS	EgoIT	rel.	ZS	EgoIT	rel.
gemma4	0.532	0.543	+2.1%	51.7	18.4	−64.4%
qwen35	0.657	0.675	+2.7%	29.0	17.4	−40.0%
qwen25vl	0.636	0.590	−7.2%	28.6	23.8	−16.8%
internvl35	0.619	0.609	−1.6%	24.1	21.7	−10.0%
internvl25	0.585	0.594	+1.5%	16.1	9.9	−38.5%
Mean	0.606	0.602	−0.5%	29.9	18.2	−33.9%

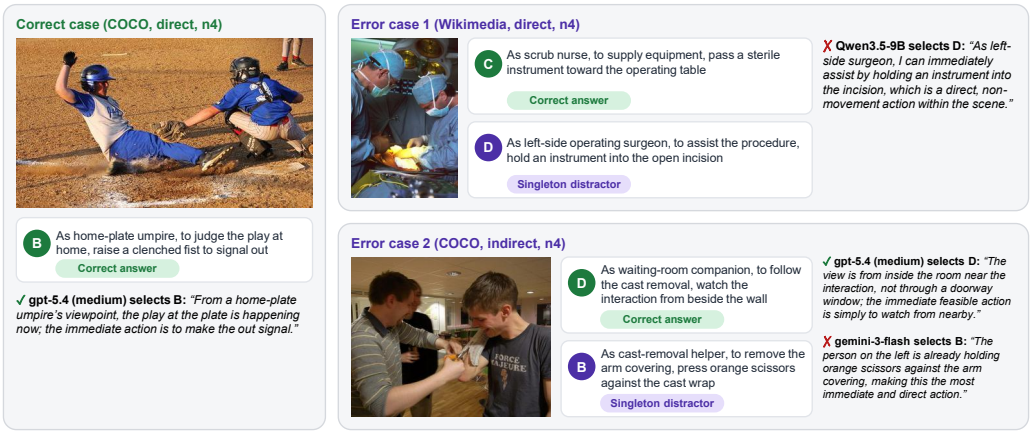

Figure 4: A correct case and two error cases. In both error cases the model selects a singleton distractor: Qwen3.5-9B on the Wikimedia set (top right) and gemini-3-flash (bottom right), shown with the correct response of GPT-5.4 (medium) under the same prompt.

on direct questions. Low accuracy is not limited to direct questions. InternVL2.5-8B scores 14.0% on indirect questions, below the chance level. With the image removed, the same model reaches 27.4% on indirect questions (Appendix I).

On n4 questions, errors concentrate on singleton distractors. Since two of the three distractors are singleton distractors, errors distributed uniformly over the distractors would fall on them 66.7% of the time. In eight of the nine configurations, the share of errors on singleton distractors exceeds this level, and the mean share over the nine configurations is 84.2%. When the evaluated models are wrong, they select other people's observed actions more often than off-vantage actions. Figure 4 shows two such errors alongside a correct case.

Humans and models also differ in how the n4 format affects them. Human accuracy falls from 99.2% on n3 to 86.8% on n4, whereas GPT-5.4 (medium) falls from 66.3% to 65.7%. The two also differ in which distractor the n4 errors select: the singleton share is 36.2% for humans and 62.6% or higher for every configuration.

4.3 Effect of Fine-tuning

Fine-tuning on EgoIT-99K. As shown in Table 2, fine-tuning on EgoIT-99K lowers EgoGapBench accuracy in all five models, with a mean relative change of −33.9%. Gemma4, for example, drops from 51.7% to 18.4%, below the chance level. In contrast, on EgoThink, the mean relative

Table 3: Fine-tuning on the EgoGapBench training split, open-source models only. Left: accuracy (%) after fine-tuning. Zero-shot values are in Table 1. Right: fraction of all n4 questions on which a singleton (Sing.) or off-vantage (Off-v.) distractor is selected (%), before (ZS) and after fine-tuning (FT).

	Accuracy (%)					Sel. rate (% of n4)			
						Sing.		Off-v.	
Model	Overall	n3	n4	Direct	Indirect	ZS	FT	ZS	FT
qwen35	75.2	83.3	62.0	63.4	97.1	69.4	14.9	6.3	23.0
gemma4	56.9	61.2	50.0	38.6	90.9	35.3	31.7	15.7	18.3
qwen25vl	54.1	58.1	47.6	40.8	78.9	63.4	34.3	9.7	18.1
internvl35	41.3	43.9	37.2	29.7	62.9	74.6	53.1	2.9	8.9
internvl25	38.6	39.3	37.4	29.2	56.0	77.7	51.3	4.5	11.3

Table 4: Effect of the explicit exclusion instruction on the five open-source models. Left: accuracy (%) with the instruction. Right: fraction of all n4 questions on which a singleton (Sing.) or off-vantage (Off-v.) distractor is selected (%) with the instruction. Subscripts give the change from the default prompt (Table 1).

	Accuracy (%)					Sel. rate (% of n4)	
Model	Overall	n3	n4	Direct	Indirect	Sing.	Off-v.
gemma4	$70.6_{+18.9}$	$77.2_{+23.8}$	$59.9_{+10.9}$	$56.2_{+22.2}$	$97.4_{+12.8}$	$7.9_{-27.4}$	$29.3_{+13.6}$
qwen35	$57.1_{+28.1}$	$62.3_{+30.4}$	$48.7_{+24.4}$	$43.7_{+23.7}$	$82.0_{+36.3}$	$33.5_{-35.9}$	$17.8_{+11.5}$
qwen25vl	$45.4_{+16.8}$	$48.4_{+18.8}$	$40.6_{+13.6}$	$35.7_{+13.9}$	$63.4_{+22.3}$	$46.9_{-16.5}$	$12.6_{+2.9}$
internvl35	$41.9_{+17.8}$	$44.0_{+18.9}$	$38.5_{+16.0}$	$34.6_{+14.1}$	$55.4_{+24.5}$	$52.4_{-22.2}$	$9.2_{+6.3}$
internvl25	$24.7_{+8.6}$	$24.3_{+9.3}$	$25.4_{+7.6}$	$24.0_{+6.8}$	$26.0_{+12.0}$	$68.3_{-9.4}$	$6.3_{+1.8}$

change is only -0.5%. The same fine-tuning can therefore affect understanding of first-person visual inputs and EAS differently.

Fine-tuning on the EAS training split. As shown in Table 3, fine-tuning on the EgoGapBench training split does raise accuracy in all five open-source models, and the ranking of the models by overall accuracy is preserved (Spearman $\rho = 0.90$ before and after fine-tuning). However, all models remain below the human baseline. The singleton distractor selection rate decreases in all five models while the off-vantage distractor selection rate increases.

4.4 Control Analyses

Prompt variation. We re-run the zero-shot evaluation with a prompt that adopts the instruction and question phrasing of EgoPlan-Bench (Chen et al., 2026), where the question is posed in the first person ("what action should I take next?") rather than addressing the model as "you", while the scene setup and output format are kept unchanged (Appendix C). Under this prompt, the ranking of the nine configurations by overall accuracy correlates with the default prompt at Spearman $\rho = 0.88$, and five of the nine configurations remain below the chance level (Appendix H).

Text-only input. We also evaluate the nine configurations with the image removed and the prompt otherwise unchanged. On direct questions, mean accuracy is 28.5% against a chance level of 30.1%, and the highest is 38.6% for GPT-5.4 (medium). On indirect questions, mean accuracy is 50.9%, since the correct answer is an indirect response while the distractors are actions of people holding singleton roles or objects, so the correct answer can often be identified from the choice text alone (Appendix I).

Explicit exclusion instruction. We also evaluate a prompt that explicitly instructs the model not to select an action that another person in the scene is performing, inserted as a single sentence at the start of the note in the default prompt (Appendix C). This instruction supplies the rule by which EgoGapBench constructs singleton distractors, and since in many everyday scenes an action performed by another person is also appropriate for the viewer, it is a hint specific to this benchmark rather than a general strategy. As shown in Table 4, overall accuracy improves in all five open-source models but remains below the human baseline. Gemma4 rises from 51.7% to 70.6%, whereas InternVL2.5-8B rises only from 16.1% to 24.7%, still below the chance level. The singleton selection rate falls in all five models while the off-vantage selection rate, which the instruction does not address, rises.

Out-of-domain evaluation. To evaluate EAS on images outside COCO, we use the out-of-domain set built from manually collected Wikimedia images with a similar construction pipeline. On this set, the open-source models again show low zero-shot accuracy, and accuracy on direct questions is lower than on indirect questions (Appendix F). The difficulty of EAS and the gap between direct and indirect questions thus also appear on this additional set.

5 Conclusion

We proposed Egocentric Action Selection (EAS) and EgoGapBench to measure third-to-first-person confusion in action selection, separately from understanding first-person visual inputs. EgoGapBench uses multi-person scenes without the camera wearer's body, and constructs the correct answer and distractors through the singleton constraint and human verification.

All evaluated MLLMs scored below the human baseline, with lower accuracy on direct than on indirect questions, and their errors concentrated on other people's observed actions. Fine-tuning on first-person instruction data lowered EgoGapBench accuracy in all five open-source models while leaving EgoThink nearly unchanged. Fine-tuning on the EgoGapBench training split raised accuracy on held-out images, but the gap to the human baseline remained.

Performance on first-person visual inputs thus does not guarantee action selection from the user's perspective, and the two should be evaluated separately. EgoGapBench provides a basis for this evaluation.

AI use statement

Among the tasks that require disclosure, we used generative AI tools to filter candidate images, to generate action options, and to generate rationales for the training split during benchmark construction, to write parts of our code, and to translate some sentences of the manuscript. We did not use generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims or write proofs, to propose hypotheses, to design our evaluation protocol, or to interpret results. The remaining tasks requiring disclosure are not applicable to this work.

Among the tasks for which disclosure is recommended, we used generative AI tools for brainstorming, literature search, drafting and editing parts of the manuscript for readability, suggesting candidate titles, and improving the design of figures. For existing benchmarks that use an LLM for scoring, we followed their original protocol.

All AI-assisted work was reviewed by the authors. Every benchmark item was inspected by the authors, all code was read and tested for correctness by the authors, and every AI-assisted sentence was reviewed individually, with citations checked against their original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

All annotation for this work was carried out by the authors. For the human baseline, one author who was not involved in the construction of the dataset and two participants outside the study answered the sampled questions. The two external participants took part voluntarily and without compensation, and no personally identifiable information was collected.

Reproducibility statement

The construction process of EgoGapBench is described in Section 3.3, with the full image-selection pipeline in Appendix D, annotation agreement in Appendix E, and the out-of-domain set in Appendix F. The experimental setup, including the evaluated models, scoring protocol, and fine-tuning settings, is described in Section 4.1. Representative prompts are given in Appendix C, and the complete set is included in the repository. Training data, hyperparameters, and the image-level train–test split are detailed in Appendix G. The benchmark data, evaluation code, and training code are available in an anonymous repository (`https://anonymous.4open.science/r/EgoGapBench-88E6`).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pp. 657–675. Springer, 2022.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *International Journal of Computer Vision*, 134(3):118, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14302, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Google. Intelligent eyewear is coming this fall. `https://blog.google/products-and-platforms/platforms/android/android-xr-io-2026/`, 2026.

Google DeepMind. Gemini 3 Flash model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf`, 2025.

Google DeepMind. Gemma 4 model card. `https://ai.google.dev/gemma/docs/core/model_card_4`, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. Viva: A benchmark for vision-grounded decision-making with human values. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2294–2311, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Zhuohang Jiang, Xu Yuan, Haohao Qu, Shanru Lin, Kanglong Liu, Wenqi Fan, and Li Qing. Superglasses: Benchmarking vision language models as intelligent agents for ai smart glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 2165–2175, June 2026.

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right'right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14257–14267. IEEE, 2025.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551. Ieee, 2017.

Yin Li, Miao Liu, and James M Rehg. 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)*, pp. 619–635, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

Meta. Meta ray-ban display: Ai glasses with an emg wristband. `https://about.fb.com/news/2025/09/meta-ray-ban-display-ai-glasses-emg-wristband/`, 2025.

OpenAI. GPT-5.4 Thinking system card. `https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf`, 2026.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

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

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. Egonormia: Benchmarking physical social norm understanding. *arXiv preprint arXiv:2502.20490*, 2025.

Yuki Sakai, Ryosuke Furuta, Juichun Yen, and Yoichi Sato. Egoinstruct: An egocentric video dataset of face-to-face instructional interactions with multi-modal llm benchmarking. *arXiv preprint arXiv:2509.22019*, 2025.

Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons Van der Sommen, et al. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4365–4374, 2024.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

Masatoshi Tateno, Alexandros Stergiou, Risa Shinoda, Yoichi Sato, and Dima Damen. Do egocentric video-language models capture both hand-and object-centric cues? *arXiv preprint arXiv:2607.08514*, 2026.

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

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28885–28900, 2025.

Ruoxuan Zhang, Qiyun Zheng, Zhiyu Zhou, Ziqi Liao, Siyu Wu, Jian-Yu Jiang-Lin, Bin Wen, Hongxia Xie, Jianlong Fu, and Wen-Huang Cheng. Mindpower: Enabling theory-of-mind reasoning in vlm-based embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29631–29641, 2026.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL `https://arxiv.org/abs/2408.05517`.

Table 5: Pilot on EgoNormia ($N = 543$). We replace the fifth answer choice of each question and report accuracy and how often the fifth choice is selected (Pick5), in %. Original: the fifth choice is kept as "none of the above." Random action: it is replaced with an action from another video. Observed action: it is replaced with the action of another person in the same scene.

	gpt-4o		gemini-3-flash	
Fifth choice	Acc.	Pick5	Acc.	Pick5
Original (none of the above)	56.7	11.0	66.9	2.6
Random action (other video)	60.6	6.4	63.7	8.7
Observed action (same scene)	42.4	35.0	32.4	51.6

A Limitations

EgoGapBench covers a specific setting in three respects. First, every scene satisfies the singleton constraint, so the actions of other people are inappropriate for the user by construction. Second, the input is a single static image with no body of the camera wearer in view, whereas real first-person visual inputs often contain video and the wearer's hands. Third, the model selects among given answer choices rather than generating an action.

B Pilot on EgoNormia

The observation in Section 1 comes from a pilot on EgoNormia (Rezaei et al., 2025), which we report here in full. EgoNormia is a multiple-choice benchmark built on first-person videos from Ego4D, selected so that multiple people take part in the activity. Its action selection subtask gives the model video frames and five candidate actions and asks for the most normatively appropriate action for the camera wearer. One candidate is the action annotated as most appropriate in the observed context, three are actions generated as most appropriate in altered versions of that context, and the fifth is "none of the above." The pilot uses this subtask only.

Construction. EgoNormia provides per-frame descriptions of each video. Using gpt-5.4-2026-03-05, we select items in which a person other than the camera wearer is present and the final-frame description refers to an action of that person, yielding 543 items. For each item we compare three versions of the fifth answer choice: the original "none of the above," a replacement with an action drawn from a random other video, and a replacement with the action of another person observed in the same scene, taken from the per-frame description. The other four choices and the correct answer are unchanged.

Results. Table 5 reports accuracy and the rate at which the fifth choice is selected (Pick5) for gpt-4o-2024-11-20 and gemini-3-flash-preview. With the random action, Pick5 is 6.4% and 8.7%, against 11.0% and 2.6% with the original fifth choice. With the observed action, Pick5 rises to 35.0% and 51.6%, and accuracy falls from 56.7% to 42.4% and from 66.9% to 32.4%.

C Prompts

We use three prompts. The default prompt addresses the model as "you" and asks for the action the model would take as the viewer. The EgoPlan-style prompt (Appendix H) keeps the first two sentences, the note, the options, and the output format of the default prompt, and replaces the task sentence with the question phrasing of EgoPlan-Bench (Chen et al., 2026), in which "I" refers to the person asking the question, that is, the viewer. The exclusion prompt (Section 4.4) is the default prompt with one sentence inserted at the start of the note. In all three prompts, {options_block} lists the answer choices with letters and {answer_letters_spec} lists the valid letters.

Default prompt.

```
Imagine that the depicted scene is happening in front
of you right now.  You are physically present at this
location and directly seeing the situation with your
own eyes.
Your task is to choose the most appropriate action you
could take next in this situation.
Note:  An appropriate action must be possible right
now, without needing to walk, run, relocate, spend
time moving, or wait for a turn.
Options:  {options_block}
Respond with compact JSON and nothing else:  {
"answer":  {answer_letters_spec}, "rationale":  "<=25
words" }
```

EgoPlan-style prompt.

```
Imagine that the depicted scene is happening in front
of you right now.  You are physically present at this
location and directly seeing the situation with your
own eyes.
Select the best answer to the following
multiple-choice question based on the image.
Considering my current observation in the image, what
action should I take next?
Note:  An appropriate action must be possible right
now, without needing to walk, run, relocate, spend
time moving, or wait for a turn.
Behavior options:  {options_block}
Respond with compact JSON and nothing else:  {
"answer":  {answer_letters_spec}, "rationale":  "<=25
words" }
```

Exclusion prompt.

```
Imagine that the depicted scene is happening in front
of you right now.  You are physically present at this
location and directly seeing the situation with your
own eyes.
Your task is to choose the most appropriate action you
could take next in this situation.
Note:  Do not choose an action that another person in
the scene is performing; instead, observe the scene
and choose the action that is appropriate for you to
take.  An appropriate action must be possible right
now, without needing to walk, run, relocate, spend
time moving, or wait for a turn.
Options:  {options_block}
Respond with compact JSON and nothing else:  {
"answer":  {answer_letters_spec}, "rationale":  "<=25
words" }
```

Response parsing. A response is scored by the letter in the `answer` field of the returned JSON, and a response without a valid option letter is counted as incorrect. Table 6 lists every configuration with at least one such response.

Table 6: Unparseable responses, counted as incorrect. Only configurations with at least one unparseable response are listed. All other configurations, including all runs on the Wikimedia set, have none. The counts include responses that decline to choose an option. n4 gives the number among n4 questions. gemini-3-flash uses low reasoning effort.

Model	Prompt	Count (%)	n4
Zero-shot, image input			
gemma4	default	2 (0.2)	0
gemma4	EgoPlan-style	1 (0.1)	0
gemini-3-flash	EgoPlan-style	1 (0.1)	1
gemma4	exclusion	62 (6.2)	11
Zero-shot, text-only input			
gemma4	default	395 (39.5)	276
gpt-4o	default	1 (0.1)	0
internvl35	default	3 (0.3)	3
Fine-tuning on EgoIT-99K, image input			
qwen25vl	default	7 (0.7)	3
internvl35	default	1 (0.1)	1
Fine-tuning on the EAS training split, image input			
gemma4	EgoPlan-style	1 (0.1)	0
internvl35	default	6 (0.6)	3

D Data Construction Pipeline

This appendix gives the per-stage counts and the models used in the pipeline summarized in Section 3.3. Stages 1–4 select images in which two people each hold a role or object that only one person can hold, and Stages 5–6 turn the selected images into questions.

Stage 1: annotation-based filter. From the 118,287 images of COCO train2017, we use the instance and keypoint annotations alone, with no model, to keep images that contain two or more clearly depicted non-crowd people. From the keypoint annotations, we then keep only images in which at least two of these people have enough visible keypoints (at least eight) for their action to be identifiable rather than being small background figures. This stage leaves 15,446 images.

Stage 2: MLLM singleton filter. We then filter for the singleton property with two vision-language models in sequence, extracting the singleton and the actions of the people that hold the singleton property. We first use Qwen3.5-35B-A3B (Qwen Team, 2026) and then gpt-5.4-mini-2026-03-17 (OpenAI, 2026), each prompted to judge whether the scene contains people in distinct, non-duplicable roles or possessions. The two stages leave 7,925 and then 2,518 images, respectively.

Stage 3: category balancing. Because COCO can be dominated by a few activities (for example, baseball), we classify each surviving image into a short activity category using Qwen3.5-35B-A3B and cap each category at 50 images. This keeps any single activity from dominating the benchmark and leaves 660 images.

Stage 4: human verification. Two annotators independently judge whether each image meets the singleton constraint, and we keep only the images they agree on, leaving 538.

Stage 5: choice generation and annotation. For each of the 538 images, gpt-5.4-2026-03-05 generates correct-answer and off-vantage candidates, and singleton distractors are rewritten from the actions extracted in Stage 2, all in the form "As [role], to [goal], [action]." Two annotators independently label every candidate (Appendix E). Of the 2,291 candidates, 1,168 are retained as correct answers and 257 as off-vantage distractors. An image is dropped when none of its correct-answer candidates is retained, which removes 25 images and leaves 513.

Stage 6: split and assembly. We split the 513 images at the image level into 440 images for the benchmark and 73 images for the training split, so that no image appears in both. Each retained correct answer becomes one question with two singleton distractors. If the image has a retained off-vantage candidate, the question also includes it as a fourth choice (n4), and otherwise the question has three choices (n3). This yields the 1,000 benchmark questions and the 168 training questions. The order of the answer choices in each question is shuffled once at assembly and fixed for all evaluations. In the benchmark, the correct answer falls on options A–C at 32.4, 34.0, and 33.7% for n3 questions and on options A–D at 25.1, 27.2, 24.9, and 22.8% for n4 questions.

E Annotation Agreement

Two annotators independently label every candidate. A correct-answer candidate receives one of three labels, valid correct answer, off-vantage, or discard, and an off-vantage candidate receives one of two labels, valid off-vantage or discard. The off-vantage label is given when the action cannot be performed from the camera's position. Only candidates on which both annotators give the same valid label enter the benchmark (Section 3.3). Candidates on which the annotators disagree are discarded rather than adjudicated, so every retained question is one that both annotators accepted independently.

For correct-answer candidates we report a three-way Cohen's κ over all three labels and a binary κ for valid correct answer versus the other two labels, which is the decision that determines whether a candidate becomes a question. On COCO (1,686 candidates) the three-way κ is 0.51 and the binary κ is 0.66, and on Wikimedia (135 candidates) they are 0.66 and 0.80. For off-vantage candidates the binary κ for valid off-vantage versus discard is 0.39 on COCO (605 candidates) and 0.42 on Wikimedia (50 candidates). We do not pool the two candidate sets because their label spaces differ.

Per-label counts. Of the 2,291 candidates, 1,168 are kept as correct answers (767 direct, 401 indirect) and 257 as off-vantage distractors; the remaining 866 are discarded. Of the discards, 523 (23% disagreement rate over all 2,291 candidates) are candidates on which the two annotators disagree and 343 are candidates both agree to remove. Each retained correct answer becomes one item, so the kept candidates yield 1,168 items before the image-level hold-out. The off-vantage candidates provide the fourth option for n4 items. The 257 off-vantage distractors consist of 230 candidates generated as off-vantage and 27 candidates generated as correct answers that both annotators labeled off-vantage, which we reuse as distractors. Of the 185 Wikimedia candidates, 107 are retained as correct answers and 26 as off-vantage distractors (21 generated as off-vantage and 5 reused from correct-answer candidates).

F Out-of-Domain Set

To test whether the results hold on images outside COCO, we manually collect 44 Wikimedia images and then apply a pipeline similar to Section 3.3, in which image filtering uses gpt-5.4-2026-03-05 alone. The set contains 107 questions (68 direct and 39 indirect, 59 n3 and 48 n4), and the chance level is 29.6%. The fine-tuned models are trained on the COCO training split only, so no Wikimedia image is used in training.

Table 7 reports the results for the five open-source models. In zero-shot, gemma4 scores 44.9% and the other four models score below chance. Zero-shot accuracy is lower than on the COCO set for all five models, and accuracy on direct questions is lower than on indirect questions in all ten rows. After fine-tuning on the EAS training split, overall accuracy rises in all five models.

All Wikimedia Commons (`https://commons.wikimedia.org`) images used in this set are under permissive licenses (CC BY, CC0, or Public Domain). We release the Commons file-page URLs together with our annotations.

Table 7: Accuracy (%) on the out-of-domain Wikimedia set (107 questions), sorted by overall accuracy within each setting. Zero-shot and after fine-tuning on the EAS training split (COCO images only).

Model	Overall	n3	n4	Direct	Indirect
Chance	29.6	33.3	25.0	29.4	29.9
Zero-shot					
gemma4	44.9	52.5	35.4	25.0	79.5
internvl35	18.7	16.9	20.8	10.3	33.3
qwen35	17.8	22.0	12.5	8.8	33.3
qwen25vl	15.9	16.9	14.6	11.8	23.1
internvl25	14.0	11.9	16.7	10.3	20.5
Fine-tuning on the EAS training split					
qwen35	65.4	78.0	50.0	50.0	92.3
gemma4	48.6	57.6	37.5	30.9	79.5
qwen25vl	44.9	49.2	39.6	38.2	56.4
internvl25	36.4	37.3	35.4	30.9	46.2
internvl35	34.6	32.2	37.5	22.1	56.4

G Training Details

The two fine-tuning analyses use an identical training recipe and differ only in the data. Fine-tuning on EgoIT-99K uses egocentric instruction data, and fine-tuning on the EAS training split uses the 168 questions built from the 73 images held out from the benchmark (Appendix D).

Data. We sample videos and images from EgoIT-99K (Yang et al., 2025), drawing from seven of its constituent sources: Charades-Ego (Sigurdsson et al., 2018), EGTEA Gaze+ (Li et al., 2018), EgoProceL (Bansal et al., 2022), EPIC-KITCHENS, HoloAssist (Wang et al., 2023), IndustReal (Schoonbeek et al., 2024), and Ego4D. Before sampling, we exclude examples with an audio input, which is outside our scope, and video clips longer than 60 seconds or with fewer than 34 frames. We then cap the sample at 800 examples per source, giving 5,272 examples in total for one epoch. These consist of 4,596 video-based examples and 676 image-based examples (Table 8). Ego4D has 472 examples after filtering. Each example is a multi-turn dialogue, so the 5,272 examples amount to roughly 16,900 question–answer turns (3.2 turns per example on average).

EAS training split. Each training question is paired with a rationale generated by gpt-5.4-2026-03-05, which is given the image, the question, the answer choices, the correct answer, and the captions and singleton annotations produced during construction. The generation prompt instructs the rationale to state that the singletons held by the visible people are taken and to justify the correct answer in the first person. The fine-tuned model receives only the image, the question, and the answer choices, in the same default prompt as evaluation (Appendix C). The training target is the final JSON output, the correct letter and a rationale of at most 25 words. The captions and singleton annotations are used only for rationale generation and are never given to the fine-tuned model. The generation prompt is as follows.

```
You see the attached image.  You are not any of the
people visible in the scene - you are simply another
person standing at the viewpoint from which the photo
was taken.
Captions:  {cap}
The visible people and the role each one occupies:
{ppl}
The situation/question (context - reason about it, do
not restate it):  {eval_prompt}
ANSWER KEY (the marked CORRECT option is the answer -
```

Table 8: Composition of the EgoIT-99K sample used for fine-tuning. Videos and images are trained jointly.

Source	Video	Image	Total
CharadesEgo	800	0	800
EGTEA	692	108	800
EgoProceL	734	66	800
EPIC-KITCHENS	730	70	800
HoloAssist	537	263	800
IndustReal	631	169	800
Ego4D	472	0	472
Total	4,596	676	5,272

```
justify it yourself, no hints given): {answer_key}
Think it through naturally in the first person, in
just a few short sentences - not a rigid checklist.
Get across, in your own words: what the scene is;
that some roles or resources here can be held by
only one person at a time (e.g. only one batter, one
catcher, one pitcher, one ball) and the people you see
already hold those, so their actions aren't yours; and
therefore the open role/action you can take right now
from where you stand.
Output: a few short first-person sentences, then
on the FINAL line output ONLY this compact JSON
(answer MUST be "{correct}"): {"answer": "{correct}",
"rationale": "<reasoning compressed to <=25 words>"}
Rules:
- Refer to people by what is visible
(role/appearance). Do NOT say "option", "A/B/C/D",
or "label".
- Do NOT mention that you were given or told the
answer.
- The final JSON's "answer" MUST be exactly {correct}.
```

Optimization. We fine-tune with ms-swift (v4.2.3) (Zhao et al., 2024) using LoRA applied across the vision encoder and the language model (target modules: all-linear; no modules frozen). We keep the framework's default LoRA learning rate of $1e{-}4$, with rank 16, $\alpha = 32$, and dropout 0.05. Training runs for one epoch with a cosine schedule, warmup ratio 0.1, an effective batch size of 16 (per-device batch 2, gradient accumulation 8), maximum sequence length 8192, and bf16 precision. All runs use seed 42.

EgoThink evaluation. We evaluate on EgoThink (Cheng et al., 2024) by running its released evaluation code as is, which uses gpt-4o as the judge, and report the mean over its 12 dimensions of each dimension's mean score. The same judge is used before and after fine-tuning.

H PROMPT VARIATION

We rerun the zero-shot evaluation under the EgoPlan-style prompt (Appendix C), which poses the question in the first person and adds standard multiple-choice phrasing while keeping the scene setup and output format unchanged. Table 9 reports the results for all nine configurations. The ranking of the nine configurations by overall accuracy correlates with the default prompt at Spearman $\rho = 0.88$. Five of the nine configurations score below the chance level of 30.2%.

Table 9: Zero-shot results under the EgoPlan-style prompt (%). Rows follow the order of Table 1 and columns are defined as in Table 1. gpt-5.4 and gemini-3-flash use low reasoning effort unless stated otherwise.

	Accuracy (%)					Sel. rate (% of n4)		Sing. share
Model	Overall	n3	n4	Direct	Indirect	Sing.	Off-v.	(% of n4 err.)
Chance	30.2	33.3	25.0	30.1	30.2	50.0	25.0	66.7
gpt-5.4 (medium)	65.8	66.5	64.7	50.9	93.4	24.3	11.0	68.9
gpt-5.4	54.9	54.9	55.0	36.5	89.1	34.8	10.2	77.3
gemma4	49.4	53.2	43.2	32.9	80.0	40.3	16.5	71.0
gemini-3-flash	39.6	42.9	34.3	28.5	60.3	55.2	10.2	84.4
qwen35	21.8	24.6	17.3	14.0	36.3	78.8	3.9	95.3
qwen25vl	27.9	28.3	27.2	22.5	38.0	63.4	9.4	87.1
gpt-4o	20.5	24.9	13.4	10.8	38.6	80.9	5.8	93.4
internvl35	28.2	29.3	26.4	17.7	47.7	67.3	6.3	91.5
internvl25	19.4	18.9	20.2	16.8	24.3	72.8	7.1	91.1

Table 10: Zero-shot accuracy with text-only input (%), in which the model receives the question and answer choices without the image. Rows follow the order of Table 1. gpt-5.4 and gemini-3-flash use low reasoning effort unless stated otherwise. †39.5% of responses are unparseable and counted as incorrect (Table 6).

Model	Overall	Direct	Indirect
Chance	30.2	30.1	30.2
gpt-5.4 (medium)	53.4	38.6	80.9
gpt-5.4	54.3	38.5	83.7
gemma4†	27.6	20.3	41.1
gemini-3-flash	23.7	20.5	29.7
qwen35	32.8	24.5	48.3
qwen25vl	40.4	33.5	53.1
gpt-4o	39.6	25.1	66.6
internvl35	24.8	23.2	27.7
internvl25	30.7	32.5	27.4
Mean	36.4	28.5	50.9

I TEXT-ONLY INPUT

To measure how much of EgoGapBench can be answered from the question and the answer choices alone, we evaluate the nine configurations with the image removed and the default prompt otherwise unchanged. Table 10 reports accuracy against the chance level. Gemma4 produces an unparseable response on 39.5% of the questions, of which 198 are exactly the response "I need an image to perform this task. Please provide the image." These responses are counted as incorrect (Table 6). On direct questions the mean is 28.5% against a chance level of 30.1%, and the highest is 38.6% for gpt-5.4 (medium). On indirect questions the mean is 50.9% against 30.2%. In an indirect question the correct answer is an indirect response such as watching, while the distractors are actions of people in singleton roles, so the correct answer can often be identified from the choice text alone. The direct questions are therefore the relevant measure for this control.