EgoProx: Evaluating MLLMs on Egocentric 3D Proximity Reasoning Across a Cognitive Hierarchy

Supplementary Material

This is the supplementary material for the paper "EgoProx: Evaluating MLLMs on Egocentric 3D Proximity Reasoning Across a Cognitive Hierarchy". We organize the content as follows.

A. Evaluation Details

General Evaluation Setup. For all evaluation processes conducted on our benchmark, we first uniformly sample each video into 8 frames and downscale each frame to a resolution of 640×640. To ensure reproducibility, unless otherwise specified, we adopt a greedy decoding strategy for all models (*i.e.*, the temperature is set to 0, and both top-p and top-k are set to 1). The multimodal input to each model is formatted as follows: *[video frames] [text prompt]*. We use a unified inference prompt to ensure a fair comparison across models. The text prompt specifies the task objective and response constraints, incorporates the question text, and includes a minimal output-format exemplar to facilitate deterministic parsing during evaluation. Additionally, we append a zero-shot reasoning prefix [5] to encourage step-by-step inference behaviors commonly observed in instruction-tuned MLLMs. The exact prompt templates used for each task category are detailed in Section J.

Human Level Performancelet. To assess human-level performance on **EgoProx**, we adopt an evaluation procedure inspired by prior benchmarking protocols such as VSI-Bench [7]. Human participants receive both the question and its corresponding video sequence simultaneously and are allowed unlimited time to provide their responses. To conduct the evaluation, we sample a representative subset of our benchmark, selecting 50 questions per task category to ensure balanced task coverage. We recruit individuals who possess basic familiarity with spatial AI and MLLMs, and we supply clear instructions along with illustrative examples. Participants may replay the video as many times as needed to ensure thorough understanding of video context before making a decision.

B. Generalization to Other Benchmark

To further examine the hypothesis introduced in Sec.5.3 that existing MLLMs possess latent spatial knowledge acquired during large-scale multimodal pretraining, yet struggle to explicitly retrieve and operationalize such knowledge for structured spatial reasoning tasks, we conduct an additional cross-benchmark evaluation. In this experiment, we assess whether instruction tuning with data generated by our **Agentic Data Engine** improves performance on VSI-Bench [7], a benchmark designed to evaluate spatial reasoning capabilities of MLLMs.

Specifically, we evaluate on the multiple-choice subset of tasks, including Relative Direction, Relative Distance, Route Planning, and Appearance Order. We adopt the same protocol used in our cross-domain analysis experiments: we select Qwen2.5-VL-7B as the base model and construct an instruction-tuned model using 800 examples drawn exclusively from the *Exploration* category, as this task type most closely aligns with the spatial reasoning patterns present in VSI-Bench. The model is fine-tuned using LoRA on top of Qwen2.5-VL-7B via the LLaMA-Factory framework. For both the base and tuned models, we adopt a consistent evaluation protocol: input videos are uniformly sampled into 8 frames, and we use the standardized inference prompt provided by VSI-Bench during testing to ensure fairness and reproducibility.

The resulting performance is summarized in Tab. 1. Notably, instruction tuning with only 800 examples yields a substantial performance gain on the VSI benchmark, despite the large domain gap between the VSI data source and egocentric video data. This observation further resonates with our earlier hypothesis that existing MLLMs possess reasonable spatial knowledge but lack the ability to effectively extract the textual representations needed to answer spatial questions. However, the improvement is smaller than the cross-domain gains observed on our own benchmark as in Tab. 3 & 4. This is because instruction-tuning data that lies in a similar visual space as the test data enables the model to more effectively learn how to leverage

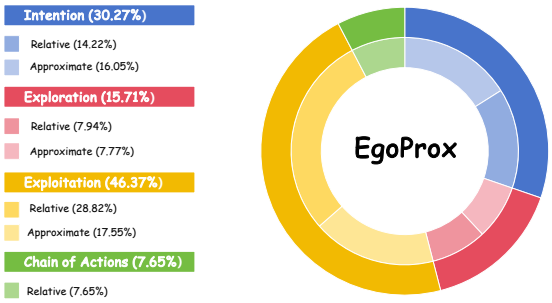

Figure 1. **Benchmark Statistics.** The distribution of tasks across four main categories in **EgoProx** with Relative and Approximate variants.

its spatial knowledge for such scenes.

C. Benchmark Statistics

EgoProx contains 2,405 VQA samples, encompassing a broad spectrum of egocentric 3D proximity reasoning tasks. These samples are derived from two complementary egocentric datasets: 1,016 from Aria Digital Twin (ADT)[3] and 1,389 from EgoExo4D[2]. Due to differences in dataset characteristics, task coverage varies across sources: *Exploration* tasks are exclusively generated from ADT, where locomotion is prominent, whereas *Chain of Actions* tasks rely solely on EgoExo4D, which contains dense, goal-oriented manipulation sequences. For the remaining task categories, samples are drawn from both datasets with balanced proportions.

As shown in Fig. 1, the benchmark is structured across four primary categories: *Intention (30.27%)*, *Exploration (15.71%)*, *Exploitation (46.37%)* and *Chain of Actions (7.65%)*, reflecting the cognitive hierarchy introduced in the main paper. Except for *Chain of Actions*, each task category includes two distinct forms of proximity measurement: *Relative* and *Approximate*.

D. Implementation details of Toolset

Formally, we define the notations in Tab. 2.

D.1. Pre-Process

For the ADT dataset, we directly obtain 3D object bounding boxes $\mathcal{O}^{3d}$, hand skeleton positions S, eye-gaze measurements E, camera poses, and egocentric video frames, and the center c_i of the objects can be calculated using the 3D bounding boxes. In contrast, the Ego-Exo4D dataset does not provide explicit 3D bounding boxes, making it difficult to localize objects in 3D space. To address this issue, we leverage the annotated interaction timestamps and approximate an object's 3D position c_i using the mean hand-skeleton position during the corresponding interaction interval. When both hands are involved, the average position of the two hand skeletons is adopted as the proxy for the object position. Furthermore, we extract keystep information from the atomic-description annotations in the Ego-Exo4D dataset to support our downstream analysis.

D.2. Toolset for 3D Analysis

Preliminary Before introducing the proposed toolset, we outline several core definitions and notations:

1. The 3D center c_i of object i is computed as

$$c_i = \left(\tfrac{1}{2}(o^{3d}_{i,1} + o^{3d}_{i,2}),\ \tfrac{1}{2}(o^{3d}_{i,3} + o^{3d}_{i,4}),\ \tfrac{1}{2}(o^{3d}_{i,5} + o^{3d}_{i,6})\right),$$

   where $o^{3d}_i \in \mathbf{R}^6$ denotes the bounding-box coordinates.
2. The camera pose is represented by the transformation matrix $T^c_s = T^d_s \times T^c_d$, where

$$T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}, \quad R \in \mathbf{R}^{3\times 3},\ t \in \mathbf{R}^3.$$

3. The camera center C corresponds to the translation component of T^c_s.
4. For angular reasoning in the world coordinate system, we discretize directions into eight canonical categories: *front*, *back*, *left*, *right*, *front-left*, *front-right*, *back-left*, and *back-right*.

Occupancy Map Generator The Occupancy Map Generator construct a navigation map $\mathcal{M}$ from the 3D bounding boxes $\mathcal{O}^{3d}$ observed in the last frame x_T to distinguish free and occupied regions for obstacle checking. Concretely, each box is projected onto the ground plane, convex hulls are computed for the projected footprints, regions enclosed by those hulls are marked as obstacles, and the interior of the outermost hull is treated as the nominal navigable area.

Exploration Path Generator Given the goal object G and the observation video $\mathcal{X}$, we can compute the center c_i of G and obtain the camera center C from the camera pose in the last frame of x_T. Then the Exploration Path Generator discretizes $\mathcal{M}$ into a 2D grid, projects the start position $p_0 = C$ and the goal position $p_K = c_i$ onto that grid, and runs an 8-connected A* search algorithm with direction-change penalties and diagonal-cut constraints to produce a feasible path. The resulting feasible path is represented as a sequence of waypoints, and each pair of adjacent waypoints defines a step $\hat{s}_i$. Note that we intentionally avoid using the actual human trajectory for navigation-step generation, as human motion exhibits high stochasticity and is difficult for MLLMs to reliably interpret.

Spatial Calculator The Spatial Calculator contains two subtools: the *Distance Calculator* and the *Direction Calculator*. The Distance Calculator projects the camera center C and object centers c_i into a unified world coordinate frame and computes Euclidean translation distances

Table 1. Cross-benchmark experimental results where best scores are colored with red. We leverage extra training data from the *Exploration* category generated by our data engine and evaluate performance across all categories in VSI-Bench [7]. The instruction-tuned model exhibits improved performance compared to the base model. Note that E, M, and H denote *Easy*, *Medium*, and *Hard* difficulty levels, respectively.

Model	Rel. Dir. (E)	Rel. Dir. (M)	Rel. Dir. (H)	Rel. Dir.	Rel. Dist.	Route Plan	Appr. Order	Avg.
Qwen2.5-VL-7B	48.39	42.33	17.16	33.99	38.17	30.41	27.18	33.21
Qwen2.5-VL-7B + Tuning	48.85	45.24	21.45	36.88	37.18	30.93	28.64	34.46

Table 2. **Summary of input notation.** For simplicity, we omit the time step for some of the notations.

Notation	Definition
$\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$	Observable video segment, where T denotes the total number for frames
$\mathcal{O}^{3d} = \{o_1^{3d}, o_2^{3d}, ..., o_N^{3d}\}$	3D object bounding boxes, where N denotes the total number of objects
E	Eye gaze data including the pose and the depth
S	Skeleton position data, including hand skeleton position
T_s^d	Transformation matrix between scene and device
T_c^d	Transformation matrix between device and camera
$\hat{h}$	Human-object interaction
$\hat{m}$	Body movement

between queried pairs (e.g., between objects i and j). The Direction Calculator computes the angle between the camera's forward direction and the vector from C to a target G, both projected onto the bird's-eye-view (BEV) plane. It first extracts the camera-plane normal from T_s^c, projects both this normal vector and the vector from C to c_i into the xOy plane, and then computes the resulting angle θ.

Gaze Parser The Gaze Parser converts 2D eye-tracking points E into 3D gaze rays in the world coordinate system. These rays differ fundamentally from the camera-plane normal. For the ADT dataset, given 3D bounding boxes o_i^{3d}, the parser checks whether the gaze ray in future frames intersects any of the six faces of o_i^{3d}, while ensuring that the corresponding object appears in the last observation frame x_T. If multiple intersections exist, the closest one to the camera center is selected. For the Ego-Exo4D dataset, the parser first selects an appropriate future frame as ground truth, inserts a marker at the eye-gaze landing position, and uses an MLLM to identify the corresponding object. Using the geometric functions above, the parser returns the intentionally interacted object (and the intersection point for ADT). If a goal object is already provided, the parser instead outputs the orientation angle required to view the object.

Affordance Detector The Affordance Detector determines whether a target object will be interacted with by the observer in future frames. It operates based on three types of $\hat{h}$, described as follows:

- When $\hat{h}$ is *afford*: For the ADT dataset, an object i is considered to be interacted with if at least one of the following criteria is satisfied: (1) its average velocity exceeds $0.05\,\mathrm{m/s}$, or (2) the hand-skeleton position from the set of skeletons S lies inside the 3D bounding box o_i^{3d}. The average velocity is computed as the translation distance divided by the time difference between the corresponding timestamps. For the Ego-Exo4D dataset, we pre-process the timestamps of annotated interaction keysteps. The Detector then checks whether future frames contain such keysteps and selects an appropriate future frame accordingly. After this determination, the Detector returns the direction and distance from the observer to the goal object in the last frame x_T of the observation segment $\mathcal{X}$, using the direction and distance computation modules described earlier.
- When $\hat{h}$ is *place*: The Detector computes the direction from the object's current center position c_i in the last observation frame x_T to its predicted position in the designated future frame. It additionally ensures that the placement location is visible within the observation video $\mathcal{X}$.
- When $\hat{h}$ is *action*: For the Ego-Exo4D dataset, the future frame is directly provided by interaction timestamps in the annotations. The Detector uses the camera pose of the last observation frame x_T and that of the future frame to compute the turn angle within the coordinate system of the camera at x_T. The final output follows the same format as described above.

Keystep Extraction Tool The Keystep Extraction Tool returns the textual keysteps in the observation video $\mathcal{X}$ including the interactive objects, the observer, and the interaction names from our pre-processed keystep data.

Chain Constructor The Chain Constructor obtains possible chains of steps and the direction between the steps. First, the Constructor obtains the processed textual keysteps from the Keystep Extraction. Then, it calculates the directions between the steps. More precisely, the *direction* is the direction between the adjcent pair of waypoints in the coordinate system of camera pose in the last frame x_T of the observation video. Regarding it as the basically correct chain, the tool provides several possible correct chains using multi-modal large language models.

D.3. Toolset Usage

In a nutshell, the 3D proximity ground truth for a given input clip sampled for each task type is constructed for each as follows:

- **Intention**: The agent invokes the Spatial Calculator to

estimate how the camera wearer adjusts head orientation toward the goal or directs gaze, as inferred by the Gaze Parser.

- **Exploration**: The agent samples a valid goal G based on visibility checks and adopts the Occupancy Map Generator and Exploration Path Generator to obtain a path composed of steps $\hat{s}$ including a series of waypoints, each providing the distance and discrete direction for exploration.
- **Exploitation**: The agent utilizes an affordance detector to identify which part of the object G the observer is grasping in the anticipation frame, where the observer will place the object G, and which direction the observer will move to interact with the object G. Which of these three types is given by $\hat{h} \in \{afford, place, action\}$ specifically.
- **Chain of Actions**: Specifically, the agent employs the Keystep Extractor to extract key action steps and their 3D spatial locations from long video segments, and to identify the key actions toward the common goal G based on future observations. It then employs an LLM to construct a set of all possible ordered combinations of key steps leading toward the same goal. Finally, The agent calls the Chain Constructor to generate a complete set of possible answers by calculating the spatial relationships among the ordered combinations of key steps.

D.4. Post-Processing

The proximity measurements include both approximate transformation and relative relationships. We discretize the transformation into intervals that are interpretable by humans. For spatial relationships, we convert the 3D directions into eight discrete orientations projected onto a specified plane. When constructing the candidate sets, we prompt the VLM to generate hard-negative distraction options. However, we provide specific instructions to ensure that these distractions do not rely on minor differences that are unsolvable even for humans.

We also conduct careful human verification to ensure both the validity (whether the questioned object is visible in the video clip and whether the positions we preprocess can approximate the real coordinates), answerability (whether the questions can be answered with the provided video clips) and accuracy (correctness of the answers) of the ground truth. For the *Chain of Actions* task, we perform a thorough examination of all possible answer sets generated by the agent. To ensure that the question-answer pairs are contextually rich, accurate, and reflective of real-world egocentric interactions, we verified the data and removed the samples that failed our quality criteria, yielding the final benchmark.

E. Additional Analysis on Experiments

In this section, we provide additional analysis of the experiments conducted on our benchmark. Among the four tasks, Chain of Action poses a particularly significant challenge to existing MLLMs, especially when compared with human performance. In addition to the inherent difficulty of multi-step reasoning over extended temporal sequences, we observe that current models, especially open-source ones, struggle with instruction following when the input context becomes substantially longer. Recall that this task requires selecting from 10 candidate actions, which further increases the burden on the model's ability to process lengthy inputs.

Regarding the other three tasks, we observe that the Exploitation task is relatively easier for both humans and models, as it requires a much shorter temporal reasoning window. Another interesting finding is that humans are markedly better at interpreting relative spatial relationships, which naturally aligns with how people describe object locations in daily life. For existing models, estimating approximate distance appears slightly easier than identifying relative spatial relationships, since the latter requires the model to correctly infer and apply an appropriate coordinate reference.

F. Training Details on Domain-specific Tuning

For all fine-tuning experiments in this work, including the cross-category experimental setting, the cross-dataset setting, and the cross-benchmark setting, we fine-tune Qwen2.5-VL-7B-Instruct with a rank-8 LoRA adapter (*target = all layers*) using the llamafactory framework. Training is performed with bfloat16 precision, AdamW optimizer, cosine learning-rate scheduling with peak learning rate 5×10^{-5}, three epochs, no warm-up, and max gradient norm 1.0. We use an effective batch size of 16 (per-device batch size of 2 with 8 gradient accumulation steps). FlashAttention is enabled automatically, and both the vision tower and multimodal projector remain frozen. All runs are trained on a single NVIDIA H20 GPU.

Cross-category fine-tuning. We fine-tune the model separately using 800 training examples per category (*Intention*, *Exploration*, and *Exploitation*) generated from our Agentic Data Engine, allowing us to assess how specialization on one reasoning type transfers across others.

Cross-dataset fine-tuning. We additionally train the model using 1,200 QA samples from each source dataset: ADT [3] and EgoExo4D [2]. This setting evaluates whether dataset-specific learning improves generalization to unseen egocentric data distributions.

Cross-benchmark fine-tuning. Finally, to test transferability to external reasoning benchmarks, we fine-tune the model using 800 training samples from the *Exploration* category only and evaluate on VSI-Bench [7] without addi-

tional adaptation.

G. Additional Visualization

We provide additional visual examples to illustrate model behaviors across different reasoning tasks in **EgoProx**. In Fig. 2, Fig. 3, and Fig. 4, we showcase cases where the intention-tuned model generates more accurate and task-aligned answers compared to the proprietary GPT-5 model across the *Intention*, *Exploration*, and *Exploitation* task categories. These examples highlight improvements in egocentric 3D Proximity reasoning after task-aware fine-tuning.

For the *Chain of Actions* setting, Fig. 5 illustrates representative model behaviors using Gemini-2.5-Pro. Unlike the other task types, which are multiple-choice, this task requires structured reasoning: the model must generate an ordered sequence of 3–5 action steps from a set of 10 candidates and additionally infer the spatial relationship between consecutive steps. This aligns with the formulation described in Sec.5.1, where an answer consists of a node sequence and corresponding spatial edges. To summarize model outcomes, we group examples into four types: fully correct (correct actions and spatial relationships), correct action sequence with spatial relationships correct under relaxed tolerance, correct action sequence but incorrect spatial relationships, and incorrect action sequence. These qualitative categories directly correspond to the quantitative metrics reported in main paper Table 2&3, namely *Act-Acc*, *Rel-Acc-S*, and *Rel-Acc-L*.

H. Limitations

A limitation of the EgoProx benchmark lies in the coverage of egocentric scenarios. Similar to most existing egocentric datasets, our current benchmark is primarily built around indoor daily activities, which means certain environments and interaction types remain underrepresented. This reflects a common bottleneck in large-scale egocentric data collection rather than a limitation of our task design. As part of future work, we plan to further diversify EgoProx by incorporating outdoor activities and other less frequent yet representative scenarios, either through new targeted data collection or through curated web-scale egocentric videos from sources such as CommonCrawl.

One limitation of our agent-based pipeline is its reliance on video metadata, such as camera pose, 3D bounding boxes, for extracting accurate 3D information. While these annotations enable precise and scalable construction of proximity ground truth, they also limit the applicability of our pipeline to datasets that provide such metadata. As future work, we plan to integrate learned 3D perception modules, for example VGGT [4], which would allow the pipeline to operate on more diverse egocentric videos without requiring pre-existing geometric annotations.

A third limitation relates to the scope of model comparisons. Following the protocol of prior Spatial AI benchmarks [7, 8], we primarily report results from prevailing general-purpose MLLMs rather than specialized spatial reasoning models. Several recent works [1, 6] have introduced architectures explicitly designed for spatial understanding, but many of these focus on generic 3D scenes or simulated environments rather than egocentric scenarios, making direct comparison less aligned with our benchmark's goals. To maintain consistency and fairness with existing evaluation practices, we therefore do not include those models in our main results. In future work, we plan to develop more advanced spatially grounded MLLMs tailored for egocentric perception and provide comprehensive comparisons against both general-purpose and spatial-specialized models on the EgoProx benchmark.

I. Reproducibility

We will release the full EgoProx benchmark, the data-generation pipeline, all evaluation scripts under the Apache License 2.0. This license permits free use, modification, and redistribution while providing explicit patent grants and protections, making the benchmark and accompanying tools suitable for both academic research and large-scale system development. These resources will enable the community to fully reproduce our results, verify the design of our agent-based pipeline, and extend the benchmark in future work.

J. Prompt Template for Evaluation

In our experiments, incorporating a chain-of-thought style prefix leads to slightly improved performance, which is consistent with findings in existing works. We further observe that providing brief examples or explicit instructions improves the parsing success rate.

For the *Intention*, *Exploration*, and *Exploitation* tasks in our **EgoProx** benchmark, we employ a unified prompt template for evaluation. In contrast, the *Chain of Actions* task differs substantially in reasoning structure and temporal planning complexity; therefore, we adopt a separate and specialized prompt template for this task. Moreover, because the *Chain of Actions* task involves varying reasoning horizons, we further provide multiple prompt variants corresponding to different action lengths.

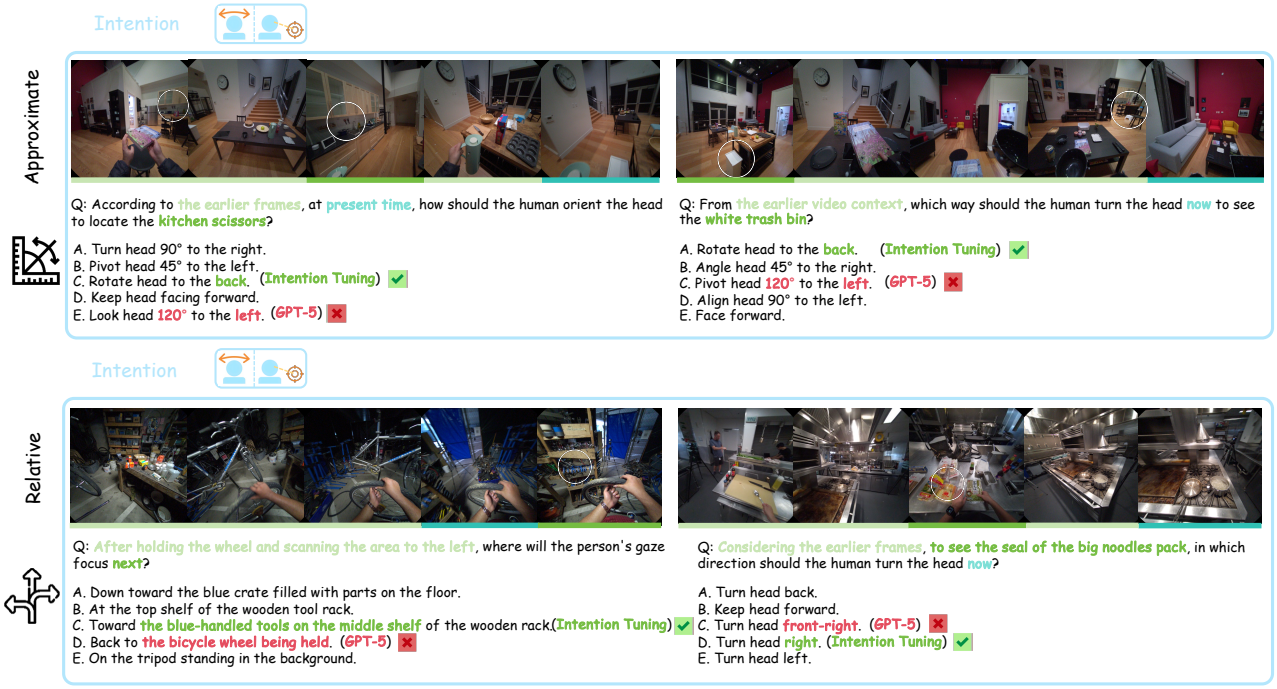

Figure 2. **Visual examples of model performance on EgoProx's *Intention* task.** We show cases where the intention-tuned model outperforms the proprietary GPT-5 model.

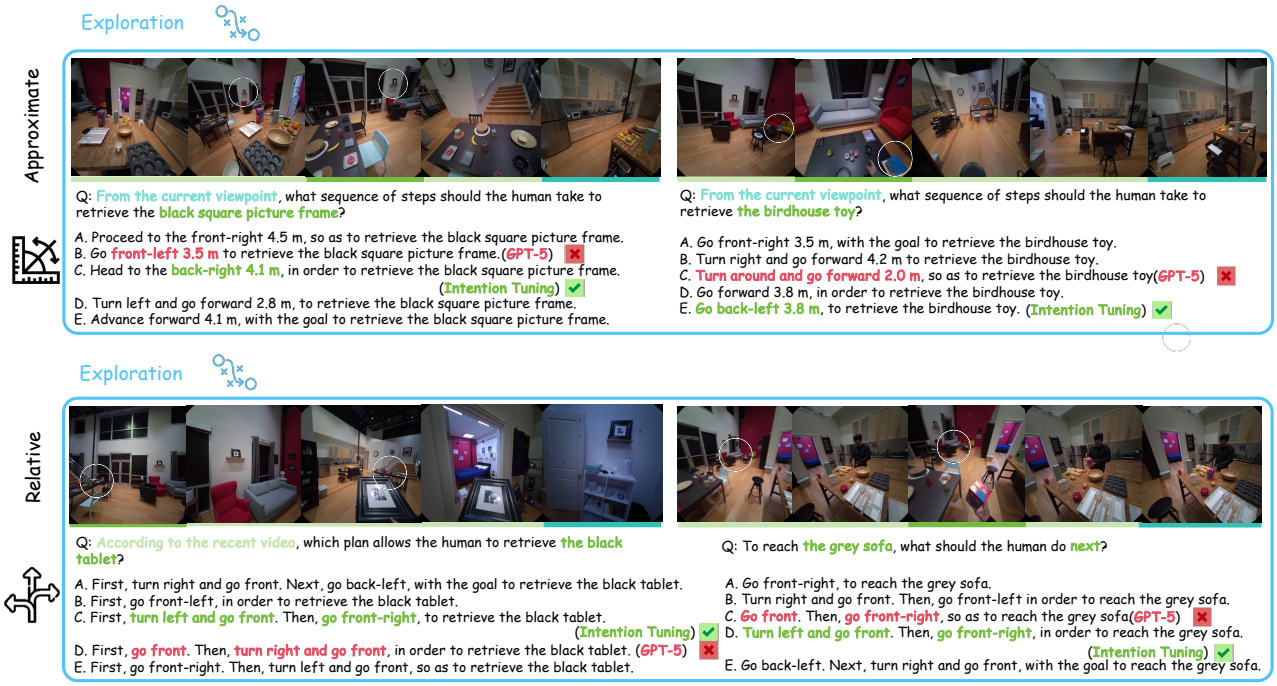

Figure 3. **Visual examples of model performance on EgoProx's *Exploration* task.** We show cases where the intention-tuned model outperforms the proprietary GPT-5 model.

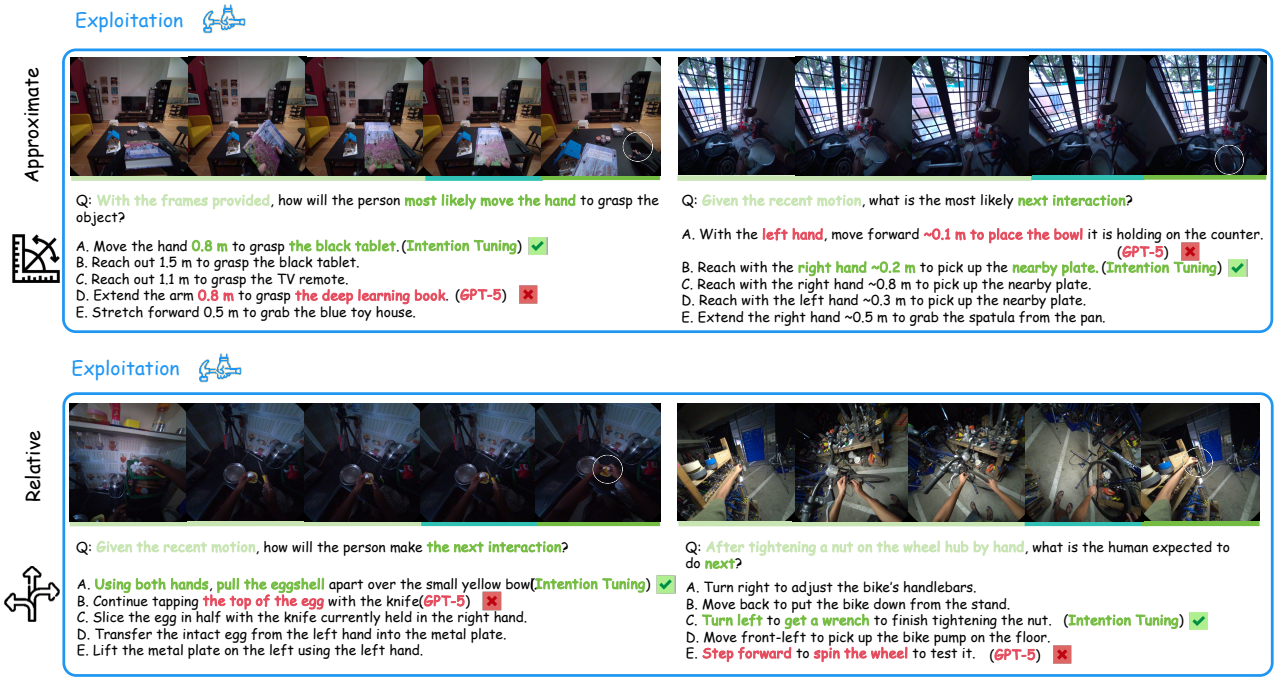

Figure 4. **Visual examples of model performance on EgoProx's *Exploitation* task.** We show cases where the intention-tuned model outperforms the proprietary GPT-5 model.

Figure 5. **Visual examples of model performance on EgoProx's *Chain of Actions* task.** We show representative cases illustrating the performance of Gemini-2.5-Pro.

Evaluation Prompt for *Intention*, *Exploration*, *Exploitation* Tasks in Egoprox (Without Chain-of-Thought)

System:
You are an expert in spatial reasoning, path planning, and human intention and behavior prediction.
You will be given a sequence of continuous first-person video frames, a question, and multiple-choice options.
The video is captured from the camera wearer's own egocentric viewpoint, meaning that "the person" or "the human" mentioned in the question refers to the camera wearer.
All spatial directions (front, back, left, right, and their diagonals) are defined in this egocentric viewpoint.
Your task is to analyze the visual content from this first-person perspective, reason about the scene in relation to the question, and select the correct answer from the provided options.

User:
[Frame 1]
[Frame 2]
[Frame 3]
...
[Frame 8]

Question: [Question text]

A. [Option A text]
B. [Option B text]
C. [Option C text]
D. [Option D text]
E. [Option E text]

Choose the most appropriate option. The selected option letter in your answer must be enclosed in angle brackets (<>).

Evaluation Prompt for *Intention*, *Exploration*, *Exploitation* Tasks in Egoprox (With Chain-of-Thought)

System:
You are an expert in spatial reasoning, path planning, and human intention and behavior prediction.
You will be given a sequence of continuous first-person video frames, a question, and multiple-choice options.
The video is captured from the camera wearer's own egocentric viewpoint, meaning that "the person" or "the human" mentioned in the question refers to the camera wearer.
All spatial directions (front, back, left, right, and their diagonals) are defined in this egocentric viewpoint.
Your task is to analyze the visual content from this first-person perspective, reason about the scene in relation to the question, and select the correct answer from the provided options.
Output format: Your final line must be: The correct answer is <>.
Example:
(Reasoning...)
The correct answer is <B>

User:
[Frame 1]
[Frame 2]
[Frame 3]
...
[Frame 8]

Question: [Question text]

A. [Option A text]
B. [Option B text]
C. [Option C text]
D. [Option D text]
E. [Option E text]

Think step by step.
Choose the most appropriate option. The option letter in your answer should be enclosed in angle brackets (<>).
Finally, end your answer with: The correct answer is <>.

Evaluation Prompt for *Chain of Actions* Task in EgoProx (Three Actions)

System:

```
You are an expert in continuous action planning and egocentric spatial reasoning.
You will receive:
(1) a short first-person video segment consisting of 8 evenly sampled frames, where the last frame is the current observation;
(2) a high-level task goal that you aim to accomplish;
(3) a set of 10 candidate keysteps, each with an integer id;
(4) a discrete set of 8 egocentric directions relative to the last frame:
  A = right, B = left, C = front, D = back, E = front-right, F = front-left, G = back-left, H = back-right.

You should regard yourself as the camera wearer, i.e., the person whose first-person viewpoint is shown in the video.
All reasoning about space, motion, and direction must be made relative to your own egocentric viewpoint as seen in the video's last frame.

Your task:
1) Choose exactly three keysteps from the candidates and order them to accomplish the goal.  Return their ids as [k1, k2, k3].
2) For each transition between consecutive keysteps (from the previous interaction to the next interaction), describe the egocentric movement direction relative to your viewpoint in the last frame.  Return these as two direction letters from {A, B, C, D, E, F, G, H} for step1→step2 and step2→step3.

All directions are defined in your egocentric frame at the last frame:  moving away from you is C (front), moving toward you is D (back), left/right are defined with respect to your viewpoint, and diagonals are E/F/G/H.

After completing your reasoning, directly output only the final answer in the following format (two lists, no extra text):
[[k1, k2, k3], [d12, d23]]

Example outputs:
[[8, 7, 3], ["F", "A"]]
[[10, 7, 9], ["E", "B"]]
```

User:

```
[Frame 1]
[Frame 2]
[Frame 3]
...
[Frame 8]

Goal:  [Goal text]

Candidate keysteps (id:  description, total = 10):
1:  [Keystep 1 text]
2:  [Keystep 2 text]
3:  [Keystep 3 text]
...
10:  [Keystep 10 text]

Egocentric direction candidates (relative to the last frame):
A: right, B: left, C: front, D: back, E: front-right, F: front-left, G: back-left, H: back-right.

Please analyze the video segment and the task goal, then provide your final answer directly in the format:
[[k1, k2, k3], [d12, d23]].
```

Evaluation Prompt for *Chain of Actions* Task in EgoProx (Four Actions)

System:

```
You are an expert in continuous action planning and egocentric spatial reasoning.
You will receive:
(1) a short first-person video segment consisting of 8 evenly sampled frames, where the last frame represents the current observation;
(2) a high-level task goal you aim to accomplish;
(3) a set of 10 candidate keysteps, each with an integer id;
(4) a discrete set of 8 egocentric directions relative to the last frame:
  A = right, B = left, C = front, D = back, E = front-right, F = front-left, G = back-left, H = back-right.

You should regard yourself as the camera wearer | the person whose first-person viewpoint is shown in the video. All reasoning about space, motion, and direction must be made relative to your own body-centered frame as seen in the last frame.

Your task:
1) Select exactly four keysteps from the candidates and order them to accomplish the goal.  Return their ids as [k1, k2, k3, k4].
2) For each transition between consecutive keysteps, describe the egocentric movement direction relative to the last frame.  Return these as [d12, d23, d34], where each direction is a single letter from {A{H}.

All directions are defined relative to your egocentric viewpoint in the last frame:  moving away from you corresponds to C (front), moving toward you corresponds to D (back), left/right are determined by your viewpoint, and diagonal movements map to E/F/G/H.

After reasoning, output only the final result in the following format (two lists, no explanation or additional text):
[[k1, k2, k3, k4], [d12, d23, d34]]

Example outputs:
[[8, 7, 3, 9], ["F", "A", "H"]]
[[10, 7, 9, 8], ["E", "B", "A"]]
```

User:

```
[Frame 1]
[Frame 2]
[Frame 3]
...
[Frame 8]

Goal:  [Goal text]

Candidate keysteps (id:  description, total = 10):
1:  [Keystep 1 text]
2:  [Keystep 2 text]
3:  [Keystep 3 text]
...
10:  [Keystep 10 text]

Egocentric direction candidates (relative to the last frame):
A: right, B: left, C: front, D: back, E: front-right, F: front-left, G: back-left, H: back-right.

Please analyze the scene and provide your final answer directly in the format:
[[k1, k2, k3, k4], [d12, d23, d34]].
```

Evaluation Prompt for *Chain of Actions* Task in EgoProx (Five Actions)

System:

```
You are an expert in continuous action planning and egocentric spatial reasoning.
You will receive:
(1) a short first-person video segment consisting of 8 evenly sampled frames, where the last frame represents the
current observation;
(2) a high-level task goal you aim to accomplish;
(3) a set of 10 candidate keysteps, each with an integer id;
(4) a discrete set of 8 egocentric directions relative to the last frame:
  A = right, B = left, C = front, D = back, E = front-right, F = front-left, G = back-left, H = back-right.

You should regard yourself as the camera wearer | the person whose first-person viewpoint is shown in the video.
All reasoning about space, motion, and direction must be made relative to your own egocentric viewpoint as seen
in the last frame.

Your task:
1) Select exactly five keysteps from the candidates and order them to accomplish the goal.  Return their ids as
[k1, k2, k3, k4, k5].
2) For each transition between consecutive keysteps, describe the egocentric movement direction relative to the
last frame.  Return these as [d12, d23, d34, d45], where each direction is a single letter from {A{H}.

All directions are defined relative to your egocentric viewpoint in the last frame:  moving away from you
corresponds to C (front), moving toward you corresponds to D (back), left/right are determined by your viewpoint,
and diagonal movements map to E/F/G/H.

After reasoning, output only the final result in the following format (two lists, no explanation or additional
text):
[[k1, k2, k3, k4, k5], [d12, d23, d34, d45]]

Example outputs:
[[8, 7, 3, 4, 5], ["F", "A", "E", "B"]]
[[10, 7, 9, 6, 8], ["E", "B", "D", "C"]]
```

User:

```
[Frame 1]
[Frame 2]
[Frame 3]
...
[Frame 8]

Goal:  [Goal text]

Candidate keysteps (id:  description, total = 10):
1:  [Keystep 1 text]
2:  [Keystep 2 text]
3:  [Keystep 3 text]
...
10:  [Keystep 10 text]

Egocentric direction candidates (relative to the last frame):
A: right, B: left, C: front, D: back, E: front-right, F: front-left, G: back-left, H: back-right.

Please analyze the scene and provide your final answer directly in this format:
[[k1, k2, k3, k4, k5], [d12, d23, d34, d45]].
```

K. Prompt Template for Training

LoRA Instruction-Tuning Prompt in EgoProx

System:
You are an expert in spatial reasoning, path planning, and human intention and behavior prediction.
You will be given a sequence of continuous first-person video frames, a question, and multiple-choice options.
The video is captured from the camera wearer's own egocentric viewpoint, meaning that "the person" or "the human" mentioned in the question refers to the camera wearer.
All spatial directions (front, back, left, right, and their diagonals) are defined in this egocentric viewpoint.
Your task is to analyze the visual content from this first-person perspective, reason about the scene in relation to the question, and select the correct answer from the provided options.

User:
[Frame 1]
[Frame 2]
[Frame 3]
...
[Frame 8]

Question: [Question text]

A. [Option A text]
B. [Option B text]
C. [Option C text]
D. [Option D text]
E. [Option E text]

Choose the most appropriate option. The option letter in your answer should be enclosed in angle brackets (<>).

Assistant:
The correct answer is <[Option Letter]>: [Chosen option text].

References

[1] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408, 2025. 5

[2] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh K. Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Romy Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, W. Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, David J. Crandall, Dima Damen, Jakob Julian Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2023. 2, 4

[3] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 2, 4

[4] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 5

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022. 1

[6] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 5

[7] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643, 2024. 1, 3, 4, 5

[8] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. *ArXiv*, abs/2505.23764, 2025. 5