EGOEXO-NEXT:BENCHMARKING VISION-LANGUAGE MODELS ON VISUAL-OPTION NEXT-STATE AND CROSS-VIEW REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly evaluated for egocentric and cross-view video reasoning, yet existing benchmarks largely focus on semantic event understanding, temporal relations, or correspondence between already observed views, leaving their ability to reason directly about future visual states underexplored. We introduce EgoExo-Next, a visual-option benchmark for dynamic visual-state reasoning, where models must identify how an observed action trajectory subsequently appears rather than predict only an action label or textual description. EgoExo-Next contains 2,503 human-curated four-choice questions from six public egocentric and ego–exo video sources and comprises four interconnected subtasks that evaluate egocentric next-state prediction, bidirectional ego–exo state correspondence, exocentric next-state prediction, and their composition in Ego-to-Exo Next-State. Extensive evaluation of proprietary, open-source, and spatial reasoning VLMs reveals a substantial human–model gap, with the best model achieving 43.81% average accuracy compared with 98.55% for humans, and the largest degradation occurring on the composed Ego-to-Exo task. These results suggest that current VLMs remain substantially limited in dynamic visual-state reasoning, particularly when temporal progression and cross-view reasoning must be composed.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated impressive reasoning capabilities across a wide range of multimodal tasks (Yin et al., 2024; Zhang et al., 2024). However, many existing evaluations still formulate reasoning through language-centered outputs, such as answering questions, describing events, or selecting among textual statements (Zheng et al., 2025; Yu et al., 2025). In physical environments, intelligent agents must also reason about how visual states evolve. Given an ongoing action, a model should be able to infer how the observed scene progresses over time, determine how the same state appears from another viewpoint, and combine these abilities to reason about a future state across perspectives (Ma et al., 2026).

Humans perform such reasoning naturally. From first-person observations of a hand reaching for a tool or manipulating an object, we can anticipate the subsequent object configuration and infer the corresponding body pose and scene layout from an external viewpoint (Ramsey et al., 2021; Gardner, 2011). This goes beyond recognizing what action is occurring: it requires reasoning about both *temporal progression*, i.e., how the observable physical configuration changes over time, and *viewpoint transformation*, i.e., how the same state is represented from another viewpoint. Their composition is particularly relevant to embodied intelligence, where an agent may perceive the environment egocentrically while demonstrations, supervision, or collaborative observations are available from exocentric views (Liu et al., 2025).

Existing benchmarks have established important foundations for egocentric and ego–exo video reasoning. EgoSchema (Mangalam et al., 2023) evaluates long-form temporal understanding in egocentric videos, EgoPlan (Qiu et al., 2026) focuses on goal-conditioned next-action planning, and EgoTempo (Plizzari et al., 2025) probes fine-grained temporal understanding over egocentric video sequences. More closely related to our setting, EgoExoBench (He et al., 2026) evaluates cross-

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

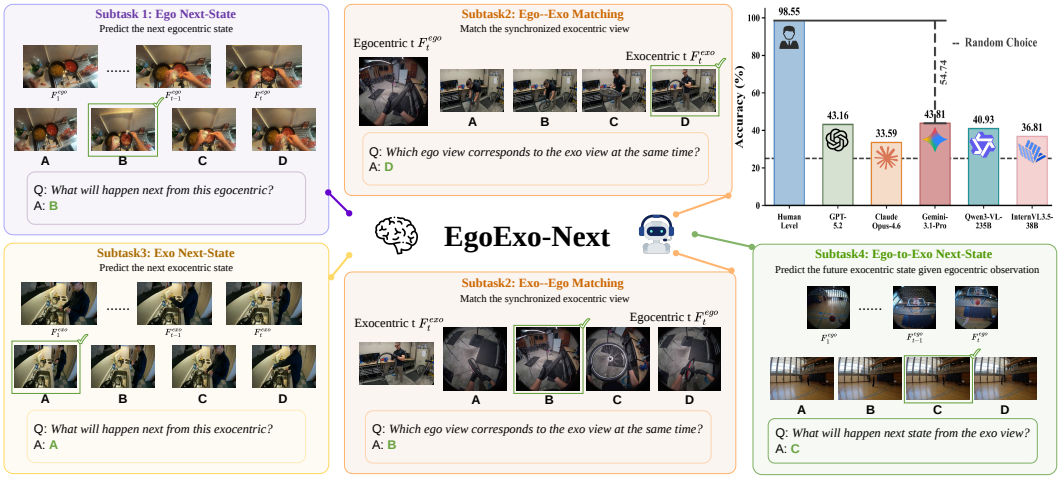


Figure 1: The EgoExo-Next benchmark. Main panels: Examples of four subtasks for next-state prediction and ego-exo view matching. Upper right: Comparison of model and human performance, showing a gap between current VLMs and humans.

view understanding through semantic alignment, viewpoint association, and temporal reasoning, including action prediction, action ordering, sequence alignment, and cross-view matching. Despite this progress, these benchmarks mainly evaluate whether models can recognize, predict, order, or align actions and events, or establish correspondence between ego–exo observations that are already given. They do not directly test whether a model can identify how the underlying visual state itself progresses into an unobserved future state.

However, reasoning about the future visual state requires a different level of understanding. Predicting what action happens next is not equivalent to identifying what the scene should look like next: an ongoing action may pass through multiple visually similar yet physically distinct states that differ in object configuration, body pose, contact relations, and spatial layout. A model may therefore correctly infer what happens next at the semantic level while still failing to identify what the scene looks like next. The distinction is even more pronounced in the ego–exo setting. Cross-view matching evaluates correspondence between already observed views, whereas EgoExo-Next requires a model to first infer an unobserved future state from preceding egocentric observations and then determine how that future state appears from an exocentric viewpoint. Thus, EgoExo-Next targets a distinct capability beyond conventional ego–exo correspondence or temporal QA: dynamic visual-state reasoning, where models identify subsequent states directly in the visual space and, in the composed setting, combine temporal progression with viewpoint transformation.

To evaluate this capability directly, we introduce **EgoExo-Next**, a visual-option benchmark in which the prediction target is the subsequent visual state itself. Rather than mapping visual evidence to an action label, textual description, or temporal relation, models must identify the target state from four temporally related image candidates. The next-state tasks further include both earlier-state and later-future distractors, so the correct answer cannot be determined simply by identifying the only future state or selecting the most temporally advanced candidate. This formulation requires models to distinguish fine-grained progression within an action trajectory, where candidate states may share the same high-level semantics but differ in object configuration, body pose, contact relations, and spatial layout. EgoExo-Next then decomposes this capability into temporal visual-state progression and cross-view state correspondence, and evaluates their composition in a more demanding setting. In particular, EgoExo-Next requires a model to infer an unobserved subsequent state from preceding egocentric observations and identify how that state appears from an exocentric viewpoint. This composition goes beyond either next-action prediction or matching already observed ego–exo content.

EgoExo-Next contains 2,503 human-curated four-choice questions constructed from six publicly available egocentric and ego–exo video sources, covering diverse indoor and outdoor scenarios. As illustrated in Fig. 1, the benchmark comprises four interconnected subtasks. **Subtask 1: Ego**

Next-State evaluates temporal visual-state progression from the egocentric perspective. **Subtask 2: Ego–Exo Matching** evaluates bidirectional correspondence between synchronized egocentric and exocentric states at the same time. **Subtask 3: Exo Next-State** evaluates temporal visual-state progression from the exocentric perspective. Finally, **Subtask 4: Ego-to-Exo Next-State** composes temporal progression with viewpoint transformation, requiring models to identify the future exocentric visual state implied by preceding egocentric observations.

We conduct extensive experiments on representative proprietary, open-source, and spatial reasoning models. Across all four subtasks, current VLMs remain substantially below human performance, with the best model achieving an average accuracy of 43.81%, compared with 98.55% for humans. The largest gap appears on **Ego-to-Exo Next-State**, where models must infer an unobserved subsequent state from preceding egocentric observations and identify how that state appears from an exocentric viewpoint. The pronounced degradation on this composed task highlights the difficulty of combining temporal visual-state progression with viewpoint transformation. Controlled analyses further show that performance depends on observation context, temporal order, and ego–exo correspondence rather than candidate-side cues alone. Together, these results establish EgoExo-Next as a challenging diagnostic benchmark for evaluating temporal visual-state progression, cross-view correspondence, and their composition.

In summary, our contributions are as follows:

- We introduce **EgoExo-Next**, a visual-option benchmark for **dynamic visual-state reasoning**, including a composed **Ego-to-Exo Next-State** task that jointly evaluates temporal progression and viewpoint transformation.
- We construct 2,503 human-curated four-choice questions from six public egocentric and ego–exo video sources, with four interconnected subtasks and a shortcut-resistant design based on non-uniform temporal sampling and controlled earlier-state and later-future distractors.
- We benchmark representative proprietary, open-source, and spatial reasoning VLMs, revealing substantial human–model gaps and showing that successful prediction depends on temporal context and cross-view evidence rather than candidate-side shortcuts alone.

2 Related Work

2.1 Egocentric VQA Benchmarks

Egocentric vision offers a natural perspective for studying human-centered perception. Recent egocentric VQA benchmarks evaluate multimodal reasoning over egocentric videos. EgoSchema targets long-form temporal understanding in egocentric streams (Mangalam et al., 2023), while EgoPlan studies goal-oriented planning from ongoing activities (Qiu et al., 2026). EgoThink evaluates first-person reasoning across multiple perceptual and cognitive dimensions (Cheng et al., 2024b). EgoTempo specifically targets fine-grained temporal understanding and is designed to require reasoning over temporal patterns rather than relying on sparse frames or commonsense priors (Plizzari et al., 2025). Earlier work such as EgoVQA also explored first-person video question answering, providing an initial testbed for egocentric VideoQA (Fan, 2019). More recently, EgoProx evaluates egocentric 3D proximity reasoning over body-object spatial relations, while EgoExoBench assesses cross-view video understanding through MCQs on semantic alignment, viewpoint association, and temporal reasoning between first- and exocentric videos (Li et al., 2026; He et al., 2026).

2.2 Ego-Exo Datasets

Paired egocentric–exocentric videos are important for studying cross-view perception and reasoning. Early datasets such as Charades-Ego (Alahari, 2018) provide wearable and external views of daily activities, while Home Action Genome (Rai et al., 2021) further introduces synchronized multi-view recordings with hierarchical activity annotations. Later datasets increasingly emphasize fine-grained procedural and interaction understanding. LEMMA (Jia et al., 2020) captures multi-agent, multi-task activities from multiple viewpoints, H2O (Kwon et al., 2021) provides synchronized RGB-D observations with detailed hand–object pose annotations, and Assembly101 (Sener et al., 2022) offers large-scale egocentric and static-camera recordings of procedural assembly activities. Recent

large-scale resources further broaden the scope of ego–exo learning. Ego-Exo4D (Grauman et al., 2024) provides synchronized first- and third-person recordings of diverse skilled activities with rich multimodal annotations, while EgoExoLearn (Huang et al., 2024) focuses on asynchronous cross-view learning between exocentric demonstrations and egocentric executions. EgoExo-Fitness (Li et al., 2024b) additionally targets full-body action understanding and cross-view skill assessment.

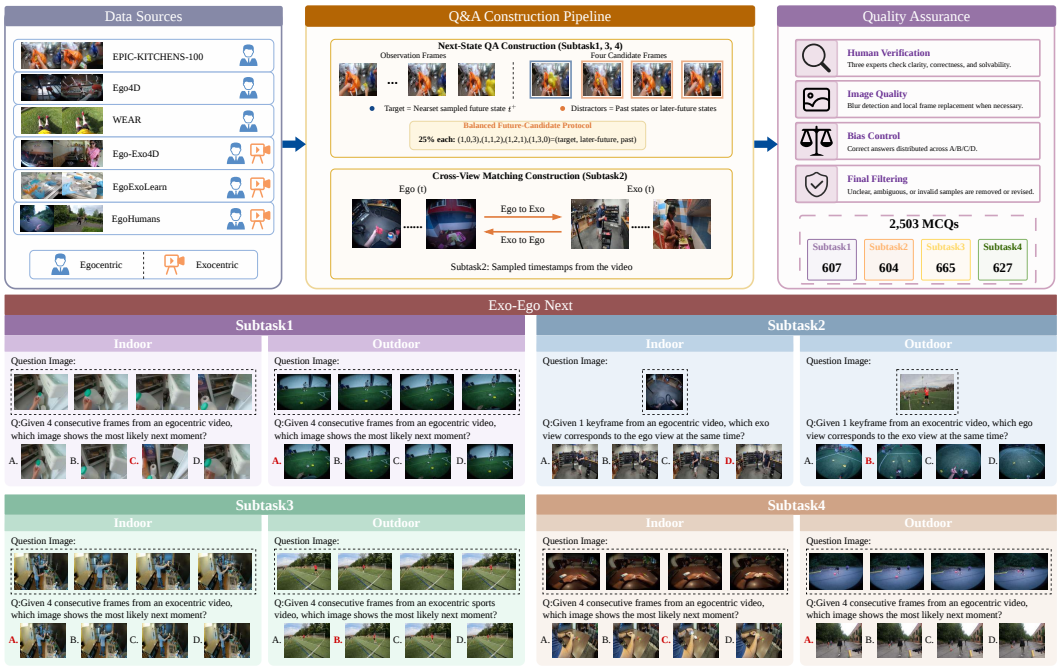


Figure 2: The upper part is the EgoExo-Next Pipeline. The lower part contains examples on EgoExo-Next. Correct answers are highlighted in red.

3 EGOEXO-NEXT

In this section, we present EgoExo-Next, a visual-option benchmark for evaluating visual-state reasoning across egocentric and exocentric perspectives. We use *visual state* to refer to the observable configuration of a scene captured in a video frame, including object configuration, human pose, contact relations, and spatial layout. The benchmark asks models to identify the appropriate target state from four visual candidates. Through four interconnected subtasks, EgoExo-Next evaluates temporal state progression, cross-view state correspondence, and their composition. We first define the four subtasks and then describe the data sources, question construction, benchmark statistics, and quality-control process.

3.1 OVERVIEW

Given egocentric or exocentric observations, EgoExo-Next requires a VLM to select the target visual state from four image candidates. Unlike benchmarks that mainly express temporal reasoning through action labels or textual descriptions, our next-state tasks directly evaluate whether models can identify how an observed trajectory subsequently appears in the visual state space.

Let F_t^{ego} and F_t^{exo} denote egocentric and exocentric visual states at time t, respectively, and let $F_{1:t}^{ego}$ and $F_{1:t}^{exo}$ denote temporally ordered observation sequences. We use $t^+ > t$ to denote the target state following the final observation; the temporal interval between observations and the target is not fixed.

Ego Next-State (Subtask 1) evaluates $F_{1:t}^{ego} \rightarrow F_{t^+}^{ego}$, measuring temporal visual-state progression from the first-person perspective. **Ego–Exo Matching (Subtask 2)** evaluates $F_t^{ego} \leftrightarrow F_t^{exo}$, measuring bidirectional correspondence between synchronized egocentric and exocentric states. **Exo**

Next-State (Subtask 3) evaluates $F^{exo}_{1:t} \rightarrow F^{exo}_{t^+}$, measuring temporal visual-state progression from the third-person perspective. **Ego-to-Exo Next-State (Subtask 4)** evaluates $F^{ego}_{1:t} \rightarrow F^{exo}_{t^+}$, requiring models to infer the subsequent state from egocentric observations and identify how it appears from an exocentric viewpoint.

The four subtasks form an interconnected diagnostic suite. Subtasks 1 and 3 evaluate temporal visual-state progression from two viewpoints, while Subtask 2 evaluates cross-view state correspondence. Subtask 4 constitutes the composed setting, requiring temporal progression from preceding egocentric observations together with cross-view reasoning about the resulting future state.

A key property of the next-state subtasks is that candidates are not restricted to earlier states. Incorrect candidates may occur either before or after the final observation. We explicitly introduce later-future distractors so that neither identifying a candidate as "future" nor selecting the most temporally advanced candidate is sufficient to determine the answer. All candidates are sampled from the same action trajectory, preserving scene and action semantics while requiring fine-grained reasoning about trajectory progression.

3.2 Benchmark Construction

As shown in Fig. 2, EgoExo-Next is constructed from human-identified action segments through a visual-option QA generation pipeline. We first collect action clips from public egocentric and ego–exo video sources, and then construct observation sequences and candidate states according to each subtask. For the next-state subtasks, temporal samples are not restricted to a fixed interval, and candidate sets contain controlled combinations of earlier states and later-future distractors. We finally apply human verification and quality filtering to ensure that each target state is visually distinguishable from the provided candidates. We next describe the data sources, question-answer construction, benchmark statistics, and quality-control process.

3.2.1 Data Source

EgoExo-Next collects images from six publicly available sources: EPIC-KITCHENS-100 (Damen et al., 2018; 2020), Ego4D (Grauman et al., 2022), WEAR (Bock et al., 2024), Ego-Exo4D (Grauman et al., 2024), EgoExoLearn (Huang et al., 2024), and EgoHumans (Khirodkar et al., 2023). These sources span diverse indoor and outdoor environments and cover cooking, object manipulation, sports, procedural activities, and human motion. EPIC-KITCHENS-100, Ego4D, and WEAR mainly provide egocentric videos for evaluating egocentric temporal visual-state progression, whereas Ego-Exo4D, EgoExoLearn, and EgoHumans provide paired or multi-view egocentric–exocentric recordings for cross-view state correspondence and composed ego-to-exo next-state reasoning. Together, these sources support the evaluation of temporal progression within each viewpoint, correspondence across viewpoints, and their composition.

3.2.2 Question-Answer Construction

To construct the visual-option QA pairs, trained annotators first inspect the raw videos and identify complete action segments, which are then extracted using Python scripts. Since action durations and dynamics vary substantially, we do not impose a fixed temporal grid or constant sampling interval. Observation frames preserve their temporal order but may have different time gaps across and within clips. For Subtasks 1, 3, and 4, each question contains four candidate states sampled from the same action trajectory. Let t denote the final observation time and $t^+ > t$ the target time. The target is the nearest sampled future state following the observations, while distractors may occur either before t or after t^+. We divide the questions equally into four groups according to the number of future candidates: 25% contain one future candidate, 25% contain two, 25% contain three, and 25% contain four. Accordingly, the four groups contain $(1, 0, 3)$, $(1, 1, 2)$, $(1, 2, 1)$, and $(1, 3, 0)$ target/later-future/past candidates, respectively. This design prevents the correct answer from being identified simply as the only future state or the most temporally advanced candidate.

Subtask 1: Ego Next-State. A temporally ordered sequence of egocentric observations is provided, and the model selects $F^{ego}_{t^+}$ from four egocentric candidates constructed using the above past/future distractor protocol.

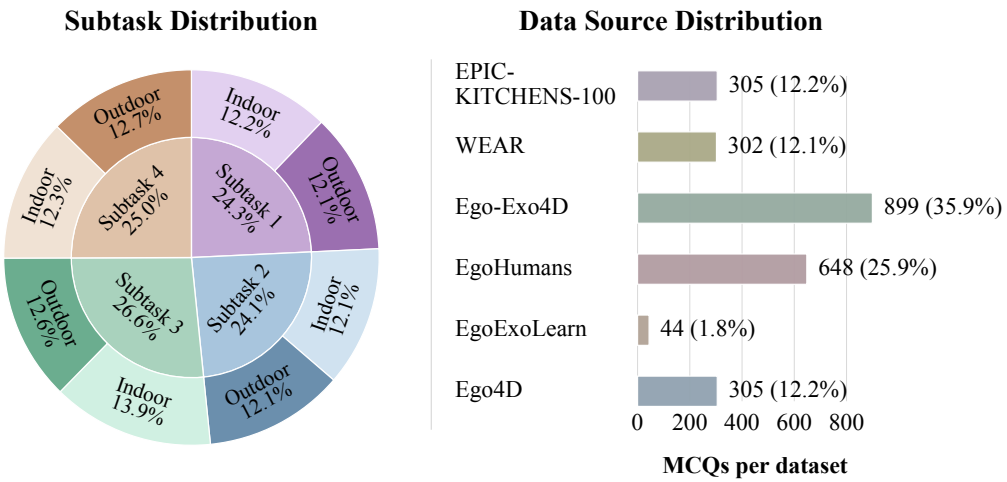


Figure 3: Benchmark composition by subtask, scene type, and data source.

Subtask 2: Ego–Exo Matching. For synchronized ego–exo videos, we sample matched timestamps and construct bidirectional ego-to-exo and exo-to-ego questions. Given a query frame from one viewpoint, the model selects its synchronized counterpart from four candidates in the other viewpoint.

Subtask 3: Exo Next-State. Subtask 3 follows the same construction as Subtask 1 in the exocentric view, requiring the model to select $F_{t^+}^{exo}$ from temporally related exocentric candidates.

Subtask 4: Ego-to-Exo Next-State. The model receives temporally ordered egocentric observations $F_{1:t}^{ego}$ and selects the corresponding future exocentric state $F_{t^+}^{exo}$. The exocentric candidates follow the same controlled past/future distractor construction, requiring both temporal progression and cross-view reasoning.

Across all subtasks, candidate order is randomized and correct-answer positions are balanced across A/B/C/D to reduce answer-position bias.

3.2.3 Benchmark Statistics

Following the construction pipeline in Fig. 2, EgoExo-Next contains 2,503 four-way MCQs collected from six public video sources. As shown in Fig. 3, Subtasks 1–4 account for 24.3%, 24.1%, 26.6%, and 25.0% of the benchmark, corresponding to 607, 604, 665, and 627 questions, respectively. Indoor and outdoor questions account for 50.5% and 49.5%, corresponding to 1,265 and 1,238 questions. For Subtasks 1, 3, and 4, questions are evenly distributed across groups containing one to four future candidates. EPIC-KITCHENS-100, WEAR, Ego-Exo4D, EgoHumans, EgoExoLearn, and Ego4D contribute 12.2%, 12.1%, 35.9%, 25.9%, 1.8%, and 12.2% of the benchmark, corresponding to 305, 302, 899, 648, 44, and 305 questions, respectively. The relatively small contribution of EgoExoLearn reflects our diversity-oriented curation: many candidate clips share similar scene and activity patterns with samples from other sources, so we retain 44 high-quality samples.

3.2.4 Quality Assurance and Filtering

To ensure benchmark quality and human-solvability, three additional reviewers, independent of the original annotators, systematically inspect every generated QA group using only the provided visual inputs. They flag samples with unclear actions, ambiguous trajectory progression, poor visual quality, indistinguishable candidate states, missing cross-view counterparts, or failed frame extraction. For Subtasks 1, 3, and 4, reviewers additionally verify that the target state is distinguishable from both earlier-state and later-future distractors based on the provided observations. A QA group is retained only when the target can be reliably identified; otherwise, it is revised or removed. We further apply blur detection and local frame replacement when necessary. Correct-answer positions are balanced across A/B/C/D, and candidate order is randomized to reduce multiple-choice bias.

Table 1: Performance on EgoExo-Next. Results are accuracy (%). Avg. denotes the mean over all eight settings. Due to the substantially higher evaluation cost of GPT-6, we evaluate it only on Subtask 4 and therefore do not report results for Subtasks 1–3 or an overall average. **Bold** and underlined values indicate the best and second-best model results, respectively, excluding human performance.

Models	Subtask 1		Subtask 2		Subtask 3		Subtask 4		Avg.
	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor	
Human Level	98.36	99.14	99.34	99.67	98.85	99.05	97.09	96.86	98.55
Proprietary Models									
GPT-6	N/A	N/A	N/A	N/A	N/A	N/A	**62.23**	**58.14**	N/A
GPT-5.2	<u>51.48</u>	**58.57**	46.53	29.31	**48.14**	<u>47.47</u>	<u>32.89</u>	30.91	<u>43.16</u>
Claude-Opus-4.6	34.75	35.71	42.39	32.12	29.23	41.14	26.33	27.04	33.59
Gemini-3.1-Pro	**54.75**	52.86	**59.44**	**42.38**	<u>44.13</u>	42.41	25.90	28.62	**43.81**
Open-Source Models									
Qwen3-VL-235B	51.15	<u>58.29</u>	<u>50.00</u>	26.82	38.40	**47.78**	26.71	28.30	40.93
Qwen3-VL-32B	46.89	43.71	47.02	26.66	37.82	34.60	25.25	<u>34.28</u>	37.03
Qwen3-VL-8B	45.18	44.65	41.44	25.59	40.17	35.24	26.75	31.13	36.27
InternVL3.5-8B	41.31	38.54	41.06	26.16	32.38	31.96	25.57	26.42	32.93
InternVL3.5-38B	45.90	44.27	46.69	28.15	36.96	37.34	25.89	29.25	36.81
LLaVA-OneVision-1.5-8B	39.02	37.58	39.74	25.50	31.23	30.70	25.24	25.79	31.85
Spatial Reasoning Models									
SpatialVLM	38.36	36.31	41.39	28.15	33.24	32.91	26.21	27.04	32.95
SpatialMLLM	41.97	40.76	44.70	30.13	35.53	36.08	26.86	28.62	35.58
RoboBrain	46.23	44.90	48.34	<u>32.45</u>	38.68	39.24	28.16	30.50	38.56

4 Evaluation on EgoExo-Next

4.1 Experiment Setup

We evaluate EgoExo-Next in a zero-shot four-way MCQ setting, where observation frames and candidate images are provided as separate visual inputs and each model outputs one option label (A–D). No in-context examples or task-specific fine-tuning are used. We evaluate representative proprietary, open-source, and spatial reasoning VLMs. Each model is evaluated three times, and we report the mean accuracy. For Subtask 2, the two matching directions are evaluated independently and averaged as $\mathrm{Acc}_{S2} = (\mathrm{Acc}_{E\to X} + \mathrm{Acc}_{X\to E})/2$. The random-guessing accuracy is 25%. We additionally evaluate training-free CLIP- and DINO-based nearest-neighbor baselines on Subtasks 1, 3, and 4 to test whether next-state selection can be explained by generic visual similarity. Detailed model configurations, deployment settings, similarity-baseline implementation, and evaluation prompts are provided in the **Appendix**.

4.2 Main Results

Table 1 yields three main observations.

(1) A substantial human–model gap remains in dynamic visual-state reasoning. Among the models evaluated across all four subtasks, Gemini-3.1-Pro achieves the highest overall accuracy of 43.81%, followed by GPT-5.2 at 43.16%, while Qwen3-VL-235B is the strongest open-source model at 40.93%. These results remain far below the human accuracy of 98.55%, revealing substantial limitations in current VLMs' dynamic visual-state reasoning across temporal progression, cross-view correspondence, and their composition. Since GPT-6 is evaluated only on Subtask 4 due to evaluation cost, it is excluded from the overall-average comparison. (2) Ego-to-Exo Next-State remains a challenging but discriminative composed reasoning task. Among models evaluated across all four subtasks, Subtask 4 exhibits the strongest performance degradation, with the best Indoor and Outdoor accuracies reaching only 32.89% and 34.28%, respectively. In contrast, GPT-6 substantially improves Subtask 4 performance to 62.23% on Indoor scenes and 58.14% on Outdoor scenes, outperforming the previous best results by 29.34 and 23.86 percentage points. This substantial improvement indicates that stronger VLMs can make meaningful progress on the composed task, suggesting that EgoExo-Next is sensitive to advances in model capability rather than merely producing floor-level performance. Nevertheless, GPT-6 remains well below human performance of 97.09% and

Table 2: Benchmark validity and shortcut analysis. Results are accuracy (%) averaged over subtasks and scene splits.

Evaluation Setting	GPT-5.2	Gemini-3.1-Pro	Qwen3-VL-235B	Qwen3-VL-32B	RoboBrain	Avg.
Overall Shortcut Test: Subtasks 1, 2, 3, 4						
Full Input	43.16	43.81	40.93	37.03	38.56	**40.70**
Candidate Only	27.25	26.75	26.00	25.80	25.75	26.31
Temporal Dependency: Subtasks 1, 3, 4						
Full Sequence	44.91	41.45	41.77	37.09	37.95	**40.63**
Last Frame Only	38.20	36.80	35.90	33.40	33.50	35.56
Shuffled Sequence	34.70	32.60	31.80	29.60	29.80	31.70
Cross-View Dependency: Subtasks 2, 4						
Full Observation	34.91	39.09	32.96	33.30	34.86	**35.02**
Mismatched Observation	26.40	27.20	25.80	25.90	26.30	26.32

Table 3: Training-free visual-similarity baselines on the three next-state subtasks. The left block reports accuracy by subtask, while the right block groups questions by the number of future candidates among the four options.

	The Type of Subtask				The Number of Future Candidates				
Method	**Subtask 1**	**Subtask 3**	**Subtask 4**	**Avg.**	**1**	**2**	**3**	**4**	**Avg.**
Random	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
CLIP	27.35	26.62	25.68	26.54	29.05	27.16	25.68	24.26	26.54
DINO	27.02	27.22	25.20	26.49	28.42	27.79	26.11	23.63	26.49

96.86%, leaving gaps of 34.86 and 38.72 percentage points on Indoor and Outdoor scenes, respectively. These results show that inferring an unobserved subsequent state from egocentric observations and identifying how that state appears from an exocentric viewpoint remains challenging even for substantially stronger models. (3) Neither scaling nor spatial specialization provides a consistent solution across the full benchmark. Proprietary models generally achieve stronger performance, although Qwen3-VL-235B remains competitive and achieves the best result on Subtask 3 Outdoor among models evaluated across all four subtasks. Scaling improves performance within some model families, such as InternVL3.5, but the trend is not consistently monotonic across Qwen3-VL variants. Specialized spatial reasoning models likewise do not consistently outperform general-purpose VLMs. Together with the substantial improvement of GPT-6 on Subtask 4, these results suggest that increased model capability can substantially improve composed visual-state reasoning, while dynamic visual-state reasoning across the broader set of temporal and cross-view settings remains far from human-level performance.

4.3 Benchmark Validity and Shortcut Analysis

We conduct controlled ablations on five representative models to test whether performance depends on the intended temporal and cross-view information rather than candidate-side cues alone. The full-input reference rows average all four subtasks (*Full Input*), temporal Subtasks 1, 3, and 4 (*Full Sequence*), or cross-view Subtasks 2 and 4 (*Full Observation*). We compare them with three controls: (1) Candidate-Side Cues: Candidate Only removes all observation images; (2) Temporal Dependency: Last Frame Only retains only the final observation, while *Shuffled Sequence* permutes the observation order; and (3) Cross-View Dependency: Mismatched Observation replaces the query with another sample from the same data source and scene split.

As shown in Table 2, *Candidate Only* reduces average accuracy from 40.70% to 26.31%, close to random guessing. Last Frame Only and *Shuffled Sequence* reduce temporal accuracy from 40.63% to 35.56% and 31.70%, respectively, while Mismatched Observation reduces cross-view accuracy from 35.02% to 26.32%. These controls do not exclude every possible low-level cue, but they show that performance depends on observation context, temporal order, and ego–exo matching rather than candidate-side information alone.

4.4 Visual-Similarity Baselines

To examine whether the next-state tasks can be solved by generic visual similarity rather than trajectory reasoning, we evaluate two training-free nearest-neighbor baselines based on CLIP and DINO. For Subtasks 1, 3, and 4, each baseline independently encodes the final observation frame and the four candidate images, computes their cosine similarities, and selects the candidate with the highest similarity. As shown in Table 3, CLIP and DINO achieve average accuracies of 26.54% and 26.49%, respectively, only slightly above the 25% random baseline. Their performance also varies across subtasks: CLIP performs slightly better on Subtasks 1 and 4, whereas DINO performs better on Subtask 3, suggesting that neither representation provides a consistently stronger similarity shortcut. We further group the questions by the number of future candidates. With only one future candidate, CLIP and DINO achieve 29.05% and 28.42%, respectively, indicating that visual similarity provides only limited cues even in the least ambiguous setting. Their performance gradually approaches chance as additional later-future distractors are introduced. With four future candidates, where every option occurs after the final observation, CLIP and DINO achieve 24.26% and 23.63%, respectively. These results indicate that generic nearest-frame similarity is insufficient for reliably identifying the target successor state. In particular, the limited advantage observed with a single future candidate disappears when multiple temporally plausible future states are included, supporting the effectiveness of the future-distractor construction in reducing similarity-based shortcuts.

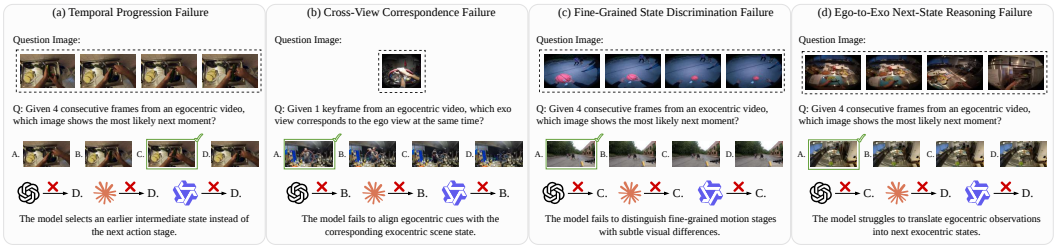


Figure 4: Representative failure cases on EgoExo-Next.

4.5 Failure Case Analysis

To further localize the limitations of current VLMs, we analyze representative failures from GPT-5.2, Claude-Opus-4.6, and Qwen3-VL-235B, covering both proprietary and open-source models. As shown in Fig. 4, we identify four recurring failure types: (1) Temporal Progression Failure: The model selects an earlier intermediate state rather than the subsequent action stage; (2) Cross-View Correspondence Failure: The model fails to align local egocentric cues, such as hand–object interactions, with the corresponding exocentric scene configuration; (3) Fine-Grained State Discrimination Failure: The model confuses visually similar candidates that differ only in subtle motion or state changes; (4) Ego-to-Exo Next-State Reasoning Failure: The model fails to combine temporal progression and viewpoint transformation to identify the next exocentric state. These cases reflect the diagnostic abilities evaluated by EgoExo-Next and show that current errors arise not only from action recognition, but also from temporal progression, cross-view correspondence, fine-grained state discrimination, and their composition.

5 Conclusion

We introduce EgoExo-Next, a visual-option benchmark for evaluating dynamic visual-state reasoning across temporal progression, cross-view correspondence, and their composition. It contains 2,503 human-curated four-way MCQs constructed from six public egocentric and ego–exo video sources and organized into four interconnected diagnostic subtasks. Extensive evaluations show that current proprietary, open-source, and spatial reasoning VLMs remain substantially below human performance, with the largest gap on the composed **Ego-to-Exo Next-State** task. Controlled analyses further show that performance depends on trajectory context, temporal organization, and ego–exo correspondence rather than candidate-side shortcuts alone. We hope EgoExo-Next supports the development of VLMs with stronger dynamic visual-state reasoning across time and perspective for embodied intelligence.

REFERENCES

Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, pp. 3–3, 2018.

Anthropic. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285/Claude%204%20System%20Card.pdf.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. Wear: An outdoor sports dataset for wearable and egocentric activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–21, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. Videgothink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623*, 2024a.

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, 2024b.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

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

Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4359–4366. IEEE, 2019.

Howard Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011.

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.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first-and third-person view video understanding in mllms. *Advances in Neural Information Processing Systems*, 38, 2026.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22072–22086, 2024.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1724–1734, 2025.

Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision*, pp. 767–786. Springer, 2020.

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19807–19819, 2023.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10118–10128. IEEE, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Jinzhao Li, Yinuo Chen, Dongxu Piao, Panwang Pan, Yifan Yu, Dong Wang, Honglei Yan, Liang Yue, Shaofei Wang, Yixin Chen, et al. Egoprox: Evaluating mllms on egocentric 3d proximity reasoning across a cognitive hierarchy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23751–23762, 2026.

Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *European conference on computer vision*, pp. 363–382. Springer, 2024b.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

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.

OpenAI. GPT-6 Astra: A New Generation of Intelligence. OpenAI, September 2026. URL `https://openai.com/index/gpt-6-astra/`.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24129–24138. IEEE, 2025.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *International Journal of Computer Vision*, 134(5):222, 2026.

Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11184–11193, 2021.

Richard Ramsey, David M Kaplan, and Emily S Cross. Watch and learn: the cognitive neuroscience of learning from others' actions. *Trends in neurosciences*, 44(6):478–491, 2021.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2022.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, et al. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *arXiv preprint arXiv:2510.25760*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

AI USE STATEMENT

In this work, we used generative AI tools only for non-core assistance, primarily for language polishing, grammatical correction, and improving the clarity and readability of the manuscript. Generative AI tools were not used to formulate the research questions, develop the proposed methodology, construct or annotate the dataset, design or conduct the experiments, analyze the experimental results, or derive the scientific claims and conclusions of this work. All core research content was developed and verified by the authors. We carefully reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work by providing detailed descriptions of the dataset construction process, task formulation, evaluation protocol, and experimental settings in the main paper and supplementary materials. In addition, we provide sample data from our dataset and the evaluation code in the supplementary materials to facilitate verification and reproduction of our benchmark evaluation. Further implementation and experimental details are also documented in the supplementary materials.

ETHICS STATEMENT

EgoExo-Next is constructed from six publicly available egocentric and ego–exo video sources. We use these data solely for research purposes and follow the access conditions and licensing requirements of the corresponding source datasets. Since the underlying videos may contain people and real-world activities, we do not use the benchmark to infer sensitive personal attributes, and any release of derived benchmark data will respect the redistribution and usage restrictions of the original sources.

Our evaluation also includes a small-scale human study involving three adult participants who independently answered the benchmark questions. We report only task-performance statistics and do not disclose personally identifying information. The benchmark is intended for evaluating visual-state reasoning in vision-language models. As with other benchmarks derived from real-world video, potential limitations include biases inherited from the source datasets, differences in scene and activity coverage, and possible privacy considerations associated with the original recordings. We encourage responsible use of the benchmark and interpretation of results within these limitations.

A APPENDIX OVERVIEW AND ORGANIZATION

This appendix provides supplementary details, benchmark comparisons, and diagnostic analyses to support and extend the main paper. The organization of the appendix is as follows:

1. **Experimental Setup (Section B):** This section provides the complete experimental protocol for EgoExo-Next, including the zero-shot four-way multiple-choice setting, the proprietary, open-source, and spatial reasoning models evaluated, the aggregation strategy for bidirectional Ego–Exo Matching, and the training-free CLIP- and DINO-based visual-similarity baselines.
2. **Evaluation Prompts (Section C):** We provide the complete evaluation prompts used for all four subtasks. Specifically, we present the unified next-state prompt for Ego Next-State, Exo Next-State, and Ego-to-Exo Next-State, as well as the bidirectional cross-view matching prompt for Ego–Exo Matching. We further describe how the prompt templates are instantiated according to observation perspective, candidate perspective, and scene setting.
3. **Comparison with Existing Egocentric and Spatial Reasoning Benchmarks (Section D):** This section compares EgoExo-Next with representative egocentric, ego–exo, and spatial reasoning benchmarks, including EgoSchema, EgoPlan-Bench, EgoThink, VidEgoThink, EgoTempo, EgoProx, and EgoExoBench. We clarify that the distinction of EgoExo-Next lies not simply in temporal or cross-view reasoning individually, but in directly iden-

tifying an unobserved subsequent visual state and, in Ego-to-Exo Next-State, composing temporal visual-state progression with viewpoint transformation within a single prediction task.

4. **Diagnosing Composed Ego-to-Exo Next-State Reasoning (Section E):** We further analyze the specific capabilities underlying the composed Ego-to-Exo Next-State task using a matched subset and controlled oracle inputs. By separately providing the current exocentric observation, the next-state egocentric observation, or both, we examine the respective contributions of cross-view correspondence, temporal visual-state progression, and their composition to the overall task difficulty.
5. **Effect of Caption Augmentation (Section F):** This section investigates whether explicit textual descriptions provide complementary semantic information for visual-state reasoning. We compare image-only evaluation with image-plus-caption evaluation using human-verified captions generated for candidate images, and analyze how explicit semantic cues affect model performance across different subtasks and scene settings.
6. **Additional Implementation Details and Results (Section G):** Finally, we provide additional implementation and evaluation details, including model deployment and inference configurations as well as participant-level human evaluation results. These details supplement the main experiments and improve the reproducibility and transparency of our evaluation.

Together, these supplementary materials provide additional methodological details, clarify the positioning of EgoExo-Next relative to existing benchmarks, and offer further evidence for the challenges of dynamic visual-state reasoning across temporal progression, cross-view correspondence, and their composition.

B Experimental Setup

EgoExo-Next is evaluated in a zero-shot four-way MCQ setting. Observation frames and candidate images are provided as separate visual inputs following Fig. 2, and each model returns one option label (A–D). No in-context examples, task-specific fine-tuning, or gradient updates are used. All models receive semantically consistent task instructions and identical output constraints.

We evaluate three model groups: (1) *Proprietary Models*, including GPT-6 (OpenAI, 2026), GPT-5.2 (Singh et al., 2025), Claude-Opus-4.6 (Anthropic, 2025), and Gemini-3.1-Pro (Team et al., 2023); (2) *Open-source Models*, including Qwen3-VL (Bai et al., 2025), InternVL3.5 (Zhu et al., 2025), and LLaVA-OneVision (Li et al., 2024a); and (3) *Spatial Reasoning Models*, including SpatialVLM (Chen et al., 2024)[1], SpatialMLLM (Wu et al., 2026)[2], and RoboBrain (Ji et al., 2025)[3].

The proprietary models and Qwen3-VL-235B are accessed through their official APIs, while the remaining models are deployed locally using their released implementations or checkpoints. Each model is evaluated three times, and the mean accuracy is reported. For Subtask 2, ego-to-exo and exo-to-ego matching are evaluated independently, and their mean accuracy is computed as $\mathrm{Acc}_{S2} = \frac{\mathrm{Acc}_{E\to X}+\mathrm{Acc}_{X\to E}}{2}$. The random-guessing accuracy for all four-way questions is 25%.

To test whether the next-state tasks can be solved through low-level visual similarity, we additionally evaluate training-free CLIP- and DINO-based nearest-neighbor baselines on Subtasks 1, 3, and 4. Each baseline independently encodes the final observation frame and the four candidate images, computes cosine similarity between the observation representation and each candidate representation, and selects the candidate with the highest similarity. These baselines use neither the preceding observation sequence nor task-specific training, providing a direct test of whether generic visual similarity alone is sufficient for next-state selection.

The complete evaluation prompts for all four subtasks, together with additional implementation and inference details, are provided in the following sections of the Appendix C.

[1] https://github.com/remyxai/VQASynth
[2] https://github.com/THU-SI/Spatial-MLLM
[3] https://github.com/FlagOpen/RoboBrain

C EVALUATION PROMPTS

C.1 EVALUATION PROMPTS FOR SUBTASK1, 3, 4

For Subtasks 1, 3, and 4, we adopt a unified next-state evaluation prompt. The prompt is instantiated according to the observation perspective and the perspective of the candidate images. Indoor and Outdoor settings differ only in the scene description and the visual cues emphasized in the instruction.

For Subtask 1, `<Observation Source>` is instantiated as "an egocentric action video," while `<Observation Perspective>` is "egocentric." The candidate image type is simply "image," and `<Candidate View Description>` is instantiated as "from the same egocentric perspective." The cross-view instruction is omitted.

For Subtask 3, the corresponding values are "an exocentric action video," "exocentric," "image," and "from the same exocentric perspective," respectively. The cross-view instruction is likewise omitted.

For Subtask 4, `<Observation Source>` is "an egocentric video," `<Observation Perspective>` is "egocentric," and `<Candidate Image Type>` is "exocentric-view image." The candidate view description is instantiated as "from an exocentric perspective," followed by the additional instruction: "The exocentric view captures the same action from an external camera angle."

C.2 EVALUATION PROMPT FOR SUBTASK 2

Subtask 2 evaluates cross-view correspondence between synchronized egocentric and exocentric observations. Each question contains one query keyframe from one perspective and four candidate images from the other perspective, and the model must identify the candidate depicting the same moment as the query. We use a unified cross-view matching prompt template for both the ego-to-exo and exo-to-ego directions.

For the ego-to-exo direction, `<Query View>` is instantiated as "egocentric," and `<Candidate View>` as "exocentric." The model is therefore asked to identify which exocentric candidate corresponds to the egocentric query image at the same time. For the exo-to-ego direction, `<Query View>` is instantiated as "exocentric," and `<Candidate View>` as "egocentric," requiring the model to identify the synchronized egocentric counterpart of the exocentric query.

For Indoor and Outdoor scenes, `<Setting>` is instantiated as "indoor" and "outdoor," respectively. In both directions, the model is instructed to compare cross-view evidence including the visible action, human pose, object configuration, spatial layout, and scene context to identify the candidate depicting the same moment from the other viewpoint.

For each direction, the query keyframe is presented before the four candidate images. In the ego-to-exo direction, the query is labeled as `Ego-centric keyframe 1:`, followed by `Exo-centric option A:` to `Exo-centric option D:`. In the exo-to-ego direction, the query is labeled as `Exo-centric keyframe 1:`, followed by `Ego-centric option A:` to `Ego-centric option D:`. The two directions are evaluated independently, and their mean accuracy is reported as the Subtask 2 result.

Template for Subtask 2

Question:
Which <Candidate View> image corresponds to the <Query View> image at the same time?
The image above is a keyframe from the <Query View> perspective in a <Setting> environment. Options A–D each show a candidate image from the <Candidate View> perspective.
Observe the query frame carefully and compare the visible action, human pose, object configuration, spatial layout, and other cross-view cues. Select the candidate image that depicts the same moment from the corresponding <Candidate View> perspective.
Only output one letter: A, B, C, or D.

Figure 5: Unified cross-view matching prompt template for the two directions of Subtask 2. The placeholders are instantiated according to the query and candidate viewpoints.

Table 4: Comparison with representative egocentric, ego–exo, and spatial reasoning benchmarks. "Visual State" indicates whether the prediction target itself is an unobserved subsequent visual state rather than an action, event, relation, or textual description.

Benchmark	Perspective	Input Type	Primary Evaluation Target	Answer Type	Cross-view	Visual State
EgoSchema	Ego	Video	Long-form temporal understanding	Text	N/A	N/A
EgoPlan-Bench	Ego	Video& Image	Goal-conditioned next-action planning	Text	N/A	N/A
EgoThink	Ego	Image	First-person perception&reasoning	Text	N/A	N/A
VidEgoThink	Ego	Video	VQA& planning& grounding& reward	Text	N/A	N/A
EgoTempo	Ego	Video	Fine-grained temporal understanding	Text	N/A	N/A
EgoProx	Ego	Video& Frames	3D proximity&spatial reasoning	Text	N/A	N/A
EgoExoBench	Ego–Exo	Paired Video	Semantic&viewpoint&temporal reasoning	Image&Text	Yes	N/A
EgoExo-Next	**Ego& Ego–Exo**	**Multi-Image**	**Visual-state progression&correspondence**	**Image**	**Yes**	**Yes**

Template for Subtask 1, 3 and 4

Question:
Given 4 temporally ordered observation frames from <Observation Source>, which <Candidate Image Type> shows the subsequent visual state following the observation sequence?
The four images above are temporally ordered observations from an <Observation Perspective> camera in a <Setting> environment. Options A–D show candidate visual states sampled from the same action trajectory <Candidate View Description>.
<Cross-View Instruction>
Observe how the action progresses across the observation frames—track <Visual Cues>—then select the candidate that depicts the subsequent state following the observation sequence.
Only output one letter: A, B, C, or D.

D Comparison with Existing Egocentric and Spatial Reasoning Benchmarks

Existing benchmarks cover several important dimensions of egocentric and embodied reasoning, but differ substantially in their prediction targets. EgoSchema (Mangalam et al., 2023) evaluates long-form temporal understanding, while EgoPlan-Bench (Qiu et al., 2026) focuses on goal-conditioned next-action planning. EgoThink (Cheng et al., 2024b) and VidEgoThink (Cheng et al., 2024a) broaden first-person evaluation to perception, question answering, planning, grounding, and reward modeling. EgoTempo (Plizzari et al., 2025) specifically targets fine-grained temporal understanding that requires integrating information across an egocentric video, whereas EgoProx (Li et al., 2026) evaluates embodied 3D proximity reasoning between the observer and surrounding objects. Although these benchmarks probe complementary aspects of egocentric intelligence, their prediction

Table 5: Diagnostic decomposition of ego-to-exo next-state reasoning on the matched subset. Results are accuracy (%).

Setting	GPT-5.2	Gemini-3.1-Pro	Qwen3-VL-235B	Qwen3-VL-32B	RoboBrain	Avg.
Original Subtask 4	32.40	28.10	27.80	30.20	29.70	29.64
+ Current Exo	48.70	44.00	45.10	38.30	40.10	43.24
+ Next-State Ego	40.20	50.10	39.40	37.70	41.30	41.74
+ Both	59.80	63.00	58.50	54.90	56.70	58.58
$\mathrm{S4} \mid \mathrm{S1}, \mathrm{S2}_{E\to X}, \mathrm{S3}$	60.50	59.20	57.40	54.10	55.80	57.40

targets are primarily actions, semantic answers, spatial relations, or other symbolic outputs rather than the subsequent whole-frame visual state itself.

EgoExoBench (He et al., 2026) is the closest benchmark to EgoExo-Next. It provides a comprehensive evaluation of first- and third-person video understanding through semantic alignment, viewpoint association, and temporal reasoning. In particular, its temporal suite includes next-action prediction, action ordering, sequence alignment, and skill evaluation, while its cross-view tasks assess semantic and spatial correspondence between ego and exo observations. Thus, EgoExoBench already evaluates both temporal reasoning and cross-view understanding. The key distinction of EgoExo-Next is therefore not the presence of either capability individually, but the *prediction target*: our next-state tasks require models to identify an unobserved subsequent visual state directly in the visual space.

This distinction is most explicit in **Ego-to-Exo Next-State**. Rather than predicting a subsequent action label or matching ego–exo content that is already observed, the model receives preceding egocentric observations and must identify the corresponding future exocentric visual state, $F^{ego}_{1:t} \rightarrow F^{exo}_{t^+}$. The task therefore composes temporal visual-state progression with viewpoint transformation within a single prediction. Moreover, the candidate set contains controlled mixtures of earlier-state and later-future distractors, preventing the target from being identified simply as the only future state or the most temporally advanced candidate. Consequently, EgoExo-Next complements existing egocentric and ego–exo benchmarks by directly evaluating **dynamic visual-state reasoning**, with Ego-to-Exo Next-State further testing its composition across time and viewpoint.

E Diagnosing Composed Ego-to-Exo Next-State Reasoning

We construct a matched subset of 200 Subtask 4 groups, with 100 Indoor and 100 Outdoor examples, and derive the corresponding Subtasks 1–3 from the same action clips and synchronized ego–exo views. We then introduce two oracle inputs: + *Current Exo* provides the synchronized exocentric frame F^{exo}_t, reducing the cross-view correspondence requirement, while + *Next-State Ego* provides $F^{\mathrm{ego}}_{t^+}$, removing egocentric next-state prediction. + *Both* provides both inputs. We also report $\mathrm{S4} \mid \mathrm{S1}, \mathrm{S2}_{E\to X}, \mathrm{S3}$, the Subtask 4 accuracy when all three diagnostic tasks are answered correctly.

Table 5 shows that the original Subtask 4 accuracy is 29.64%. (1) Cross-view correspondence and egocentric temporal progression both contribute to the task difficulty. Providing + *Current Exo* and + *Next-State Ego* increases accuracy to 43.24% and 41.74%, respectively. (2) Integrating these component abilities remains challenging. Providing both oracle inputs raises accuracy to 58.58%, while Subtask 4 accuracy reaches 57.40% on groups where all three matched diagnostic tasks are answered correctly. These results support the proposed diagnostic decomposition, while the remaining errors indicate that combining the component abilities within a unified egocentric-to-exocentric next-state reasoning task remains difficult.

F Effect of Caption Augmentation

We investigate whether explicit textual descriptions can provide complementary semantic cues for visual-state reasoning. We sample 400 questions, with 50 examples from each Indoor and Outdoor split of the four subtasks, and evaluate models under two settings: *Image Only* and *Image + Caption*.

For each question, Qwen2.5-VL-72B independently generates a concise caption for each of the four candidate images using the template shown in Fig. 7. No captions are generated for the observation

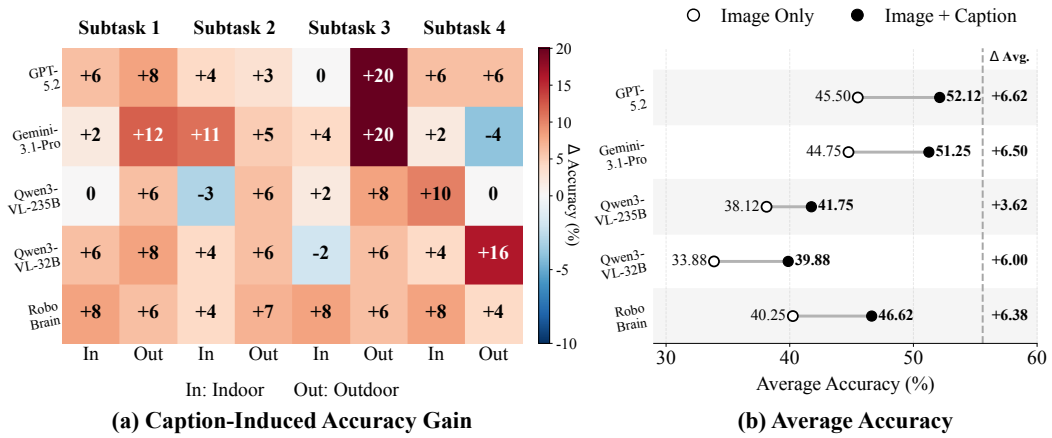


Figure 6: Effect of caption augmentation on EgoExo-Next. Left: accuracy gains across eight evaluation settings. Right: average accuracy under image-only and image-plus-caption inputs.

Caption Template

A `<Perspective>` view of `<Scene>` in a `<Setting>`. The image shows `<Situation>` involving `<Objects>`, with `<Background>` visible in the scene.

Figure 7: Prompt template used by Qwen2.5-VL-72B to generate captions for candidate images.

images, and Qwen2.5-VL-72B is excluded from the evaluated models. Human reviewers verify that each caption faithfully describes the visible perspective, scene, action or situation, objects, and background without introducing unsupported information; inaccurate descriptions are manually corrected. In the *Image + Caption* setting, each caption is placed immediately after its corresponding candidate image and option label, while the observation inputs, candidate images, and evaluation prompts remain otherwise unchanged.

As shown in Fig. 6, caption augmentation increases the overall accuracy from 40.50% to 46.33%, corresponding to a gain of 5.83 percentage points. Across 40 model–setting pairs, performance improves in 34 cases, remains unchanged in three, and decreases slightly in three. The largest average gain is observed on Subtask 3 Outdoor, where accuracy improves by 12.00 percentage points.

These results indicate that explicit semantic descriptions can complement visual information and help models distinguish candidate states. However, since the captions are generated directly from the candidate images, the improvement should be interpreted as evidence for the benefit of explicit semantic cues rather than as a direct measure of language bias.

G Additional Implementation Details and Results

G.1 Model and Inference Configurations

The proprietary models and Qwen3-VL-235B are evaluated through their official APIs, while the remaining open-source models are deployed from their official checkpoints on eight NVIDIA GeForce RTX 4090 GPUs. All evaluations are conducted in the zero-shot setting without in-context examples, task-specific fine-tuning, or gradient updates. For API-based evaluation, the temperature is set to 0, and the maximum response length is limited to 32 tokens. Each model is evaluated in three independent runs, and the mean accuracy is reported. The same evaluation instructions and output constraints are used across models.

Table 6: Participant-level human performance on EgoExo-Next. Each entry is computed from the participant's number of correct answers in the corresponding evaluation split.

Participant	Subtask 1		Subtask 2		Subtask 3		Subtask 4		Avg.
	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor	
Participant 1	97.05	99.66	99.67	99.34	99.71	98.42	96.12	97.48	98.43
Participant 2	99.67	98.34	98.68	100.00	97.99	99.68	98.06	96.23	98.58
Participant 3	98.36	99.41	99.67	99.67	98.85	99.05	97.09	96.86	98.62
Mean	98.36	99.14	99.34	99.67	98.85	99.05	97.09	96.86	98.55

G.2 Human Evaluation Details

Human performance is evaluated by three adult participants who were not involved in benchmark annotation or quality control. Each participant independently completes all 2,503 questions using the same observation and candidate images provided to the evaluated models. Candidate order is randomized, and participants are given access to all input images without a strict time limit. Table 6 reports the participant-level accuracy for each Indoor and Outdoor split. Human performance in each setting is computed as the mean of the three individual accuracies. Following the main results, *Avg.* denotes the unweighted mean over the eight subtask scene settings.