ROBOCHRONO: A REAL ROBOT BENCHMARK FOR STREAMING TASK UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding ongoing robot manipulation requires models to interpret visual observations in relation to interaction history and task progress. We introduce RoboChrono, a benchmark for streaming task understanding comprising 39 scenarios and 34,713 evaluation instances, constructed from real robot executions and complementary bare-hand human recordings. The benchmark evaluates seven tasks grouped into recognition, alignment, and temporal grounding, covering action understanding and anticipation, visual correspondence, temporal ordering, and action localization. Zero-shot evaluation of 18 vision-language models reveals substantial differences across tasks. GPT-6-Astra achieves 98.3% accuracy on Frame Matching but 68.3% on Frame Ordering, while RynnBrain1.1-122B-A10B exhibits a larger gap, reaching 95.4% and 32.9%, respectively. Input ablations on matched questions with five open-weight models further reveal distinct dependencies on visual evidence: removing visual observations reduces Current Action Recognition accuracy by 22.1 percentage points, whereas Next Action Prediction decreases by only 0.7 points. These findings show that strong visual matching does not consistently coincide with strong temporal ordering, and suggest that next-action prediction can be supported by task and action priors even when visual evidence is unavailable. RoboChrono provides a diagnostic setting for examining these differences, highlighting the need for capability-specific evaluation beyond aggregate scores when assessing task understanding in robot manipulation.

1 INTRODUCTION

Understanding an ongoing robot manipulation task requires more than recognizing what is visible in the current frame. An embodied agent must reason over the preceding interaction history to infer the current task state, distinguish completed from ongoing actions, and anticipate what may happen next. Prior work on egocentric video has similarly shown the importance of temporal context for fine-grained action understanding and anticipation (Damen et al., 2018; Furnari & Farinella, 2019; Jia et al., 2022). The problem is particularly challenging from robot-mounted viewpoints, where occlusion, motion, and partial observability can limit the evidence available from any single observation or camera (Jangir et al., 2022; Khazatsky et al., 2024). We study streaming temporal task understanding: interpreting the progress of robot manipulation using only visual observations available up to the queried moment, without access to any subsequent frames or evidence.

Existing benchmarks have advanced several complementary aspects of embodied visual reasoning. Human egocentric benchmarks evaluate temporal and causal reasoning over recorded activities (Jia et al., 2022), while video based spatial benchmarks focus on understanding spatial relationships from visual observations (Yang et al., 2025). Robot centric datasets and benchmarks further connect visual understanding with long horizon reasoning, planning, and physical interaction (Sermanet et al., 2024; Chen et al., 2026a; Bao et al., 2026), and recent embodied foundation model benchmarks broaden evaluation toward multimodal reasoning in physical environments (Ji et al., 2025; Dang et al., 2026). However, these settings largely emphasize reasoning over completed or retrospectively available observations, where later frames may provide evidence for interpreting earlier events. In an ongoing robot execution, such future information is unavailable, and models must infer the current task state, distinguish completed from ongoing actions, recover temporal structure, and anticipate subsequent behavior from the observation history available at the queried moment. This creates a distinct evaluation challenge that is not fully captured by conventional video question answering

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

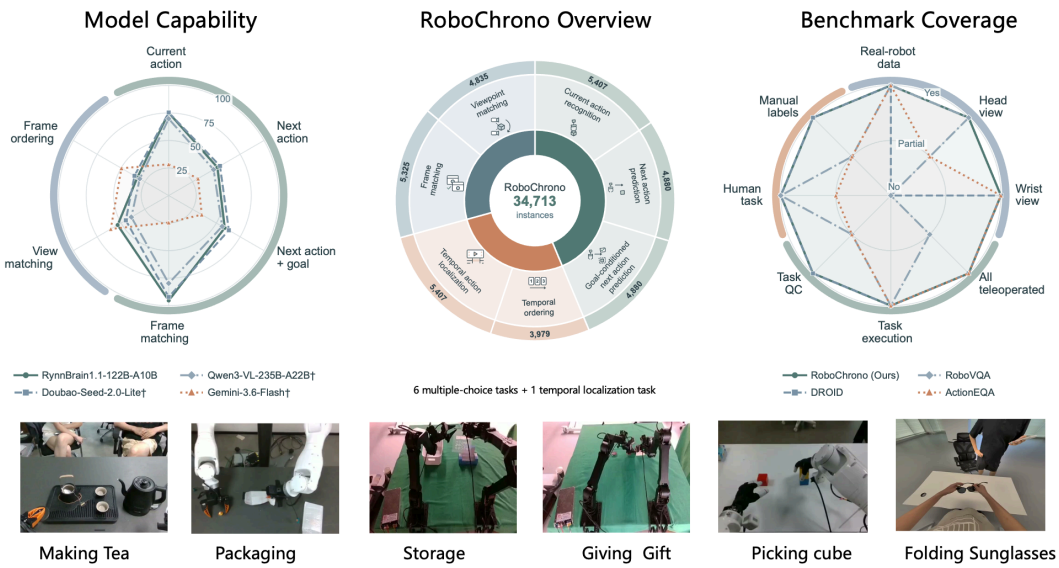

Figure 1: Overview of RoboChrono. Left: Performance profiles of four representative multimodal models across the six multiple-choice evaluation tasks. Center: Distribution of 34,713 evaluation instances across seven tasks. Right: Comparison of data collection and annotation attributes across representative robot benchmarks and datasets.

or full video reasoning. We therefore focus on streaming temporal task understanding and ask whether multimodal models can reliably interpret the temporal progress of real robot manipulation under explicitly restricted observation histories. Table 1 compares RoboChrono with representative egocentric and robot-centric datasets and benchmarks across data sources and annotation types.

We introduce **RoboChrono**, a real-world robot benchmark for streaming temporal task understanding. RoboChrono contains 39 manipulation scenarios, 1,816 successful execution videos, approximately 25 hours of real robot interaction data, and 34,713 evaluation instances. As illustrated in Figure 1, the benchmark evaluates seven tasks covering action interpretation and anticipation, visual evidence association, and temporal organization. In the streaming setting, models can only access observations available up to the queried moment and cannot use future visual evidence. We evaluate 18 open-source and commercial multimodal models across all 39 scenarios under a zero-shot protocol. Our results reveal substantial differences in capability profiles, particularly on temporal ordering, localization, and future action reasoning, which aggregate scores often obscure.

Our contributions are threefold. First, we introduce **RoboChrono**, a large-scale real robot benchmark specifically designed for streaming temporal task understanding. Second, we establish a unified evaluation framework with seven complementary tasks that assess action understanding, anticipation, cross-view and temporal association, ordering, and temporal localization under controlled observation histories. Third, we provide a systematic zero-shot evaluation of 18 multimodal models across all 39 scenarios, revealing persistent limitations in understanding the temporal structure and progression of real-world robot manipulation, particularly in ordering and localization.

2 Related Work

Egocentric Video Understanding. Egocentric video has become an important setting for studying temporal understanding from the first person perspective. Large scale datasets such as Ego4D (Grauman et al., 2022) and Ego Exo4D (Grauman et al., 2024) support research on perception, episodic memory, action recognition, and forecasting. Recent benchmarks further emphasize higher level task reasoning. EgoTaskQA (Jia et al., 2022) evaluates temporal and causal reasoning over actions and state changes, while EgoSchema (Mangalam et al., 2023), EgoPlan Bench (Chen et al., 2026b), and VidEgoThink (Cheng et al., 2024) study long-video question answering, task planning,

Table 1: Comparison of egocentric and embodied robot task benchmarks and datasets. We compare data sources, robot-mounted views, task collection and annotation protocols, and evaluation scale. RoboChrono uses real-robot executions with complementary onboard views and human-defined temporal annotations, evaluated under restricted observation histories.

Benchmark / Dataset	Primary focus	Data source	Robot-mounted view		Task collection			Curation		Evaluation scale
		Real robot	Head	Wrist	All teleop	Task exec.	Task QC	Human task	Manual labels	
(i) Human-worn or scene-centric embodied cognition benchmarks										
EgoTaskQA Jia et al. (2022)	Human task reasoning	✗	✗	✗	✗	✓	△	✓	△	2k videos / 40k QA
VSI-Bench Yang et al. (2025)	Spatial cognition	✗	✗	✗	✗	✗	✗	△	△	288 videos / 5k+ QA
RoboSpatial Song et al. (2025)	Spatial / affordance QA	✗	✗	✗	✗	✗	✗	✗	✗	1M images / 3M QA
ERQA Team et al. (2025)	Embodied reasoning QA	△	△	△	✗	✗	✗	✓	✓	400 QA
MMSI-Bench Yang et al. (2026)	Multi-image spatial QA	△	△	△	✗	✗	✗	✓	✓	1,000 QA
RynnBrain-Bench Dang et al. (2026)	Cognition + localization	✗	✗	✗	✗	✗	✗	△	△	3,616 clips / 12k QA
(ii) Physical-robot embodied benchmarks and demonstration resources										
JRDB Martin-Martin et al. (2021)	Perception / tracking	✓	✓	✗	✗	✗	✗	✗	✓	64 min / 3.5k tracks
RoboVQA Sermanet et al. (2024)	QA + long-horizon planning	✓	✓	✗	△	✓	△	✓	△	238 h / 829.5k pairs
DROID† Khazatsky et al. (2024)	Robot demonstrations	✓	✗	✓	✓	✓	✓	✓	✓	76k traj. / 350 h / 86 tasks
Open-X VQA / Robo2VLM-1 Chen et al. (2026a)	Robot interaction QA	✓	△	△	✓	✓	△	△	✗	176k traj. / 684.7k QA
ShareRobot (Aff./Traj.) Ji et al. (2025)	Planning / affordance / trajectory	✓	△	△	△	✓	✓	△	△	51.4k inst. / 1.028M QA
ActionEQA Bao et al. (2026)	Action-state reasoning	✓	△	✓	✓	✓	△	△	△	2,629 eps. / 8,795 QA
RoboChrono (Ours)	**Robot-ego temporal understanding**	✓	✓	✓	✓	✓	✓	✓	✓	**39 scenarios / 34,713 QA**

Definitions. ✓: yes; ✗: no; △: partial or mixed. "Real robot" denotes physical robot data; "Head" and "Wrist" denote global and end-effector-mounted views, respectively. "All teleop" and "Manual labels" indicate fully human-teleoperated executions and human-produced annotations, respectively. "Task QC" denotes execution quality control.

and embodied reasoning. However, these benchmarks are primarily built from human-worn videos, whose viewpoints and motion patterns differ from those of cameras mounted on physical robots.

Robot-Centric Embodied Understanding. Recent work has extended visual reasoning from human activities to robot interaction data. RoboVQA (Sermanet et al., 2024) constructs large-scale video question answering data from both human and robot embodiments to study long-horizon robotic reasoning. Robo2VLM (Chen et al., 2026a) converts real robot trajectories into visual question answering samples by extracting manipulation phases and supervision from robot states, end-effector poses, gripper measurements, and force signals. ActionEQA (Bao et al., 2026) evaluates the relationship between semantic actions and physical state transitions. More recent embodied benchmarks, including RoboBench (Luo et al., 2025), RoboBrain (Ji et al., 2025), and RynnBrain (Dang et al., 2026), further broaden evaluation toward embodied perception, spatial reasoning, affordance understanding, and planning. These efforts demonstrate the importance of evaluating multimodal models on physically grounded robot data, although their data sources, embodiments, camera configurations, and annotation pipelines are often heterogeneous or constructed for broad embodied capabilities rather than controlled temporal task understanding in streaming settings.

Streaming Video Understanding. Most video understanding benchmarks assume that the complete video is available before answering a query. This offline setting differs from embodied agents that must interpret observations while an interaction is still unfolding. StreamingBench (Lin et al., 2026) evaluates multimodal models with questions presented at different timestamps, targeting real-time perception and accumulated context. OVBench (Huang et al., 2024) further organizes online video understanding around perception, memory, and prediction over past, present, and future information. Recent work also reveals a tradeoff between immediate perception and long-range memory (Shen et al., 2026). However, existing streaming benchmarks mainly use general videos and human-centered activities rather than physical robot manipulation with explicit task structure.

Different from existing egocentric, robotic, and streaming benchmarks, RoboChrono focuses on streaming task understanding from controlled real robot executions. All episodes are collected under a unified teleoperation and quality control protocol with complementary head and wrist observations. The benchmark covers both gripper and dexterous-hand manipulation and provides human-annotated temporal events for every execution. These annotations support evaluation of task-stage understanding, event localization, question answering, and next-action prediction at meaningful moments during execution. To our knowledge, RoboChrono is the first benchmark to jointly evaluate these properties in a unified real-robot setting under restricted observation histories.



Figure 2: Overview of the task coverage in RoboChrono. The benchmark includes diverse manipulation objects and action primitives drawn from 39 real-world manipulation scenarios.

3 Benchmark Design and Evaluation

3.1 Benchmark Overview

RoboChrono evaluates whether multimodal models can understand the progress of real-world robot manipulation from progressively revealed visual observations. The benchmark focuses on three capabilities: understanding ongoing actions and anticipating subsequent behavior, establishing correspondences across observations and camera views, and localizing task-relevant events in time. At an evaluation time t, a model receives the observed video prefix $\mathcal{O}_{\leq t}$ and a task-specific query q_t. It must answer using the available observations without accessing future frames. Depending on the evaluation task, the required output is either a multiple-choice answer or a temporal interval. Explicit task goals are provided for goal-conditioned next-action prediction. The benchmark comprises seven evaluation tasks constructed from temporally annotated manipulation executions. Figure 2 illustrates the range of manipulation objects and action primitives covered by RoboChrono, while Figure 3 summarizes the complete benchmark construction and evaluation pipeline, from frame-accurate temporal annotation to VQA generation and unified zero-shot model evaluation.

3.2 Data Collection and Temporal Annotation

We design 39 manipulation tasks, including 29 gripper based tasks, 5 dexterous hand tasks, and 5 ego hand tasks. The task set spans object grasping, picking, placing, transferring, stacking, packaging, cleaning, and tool use, covering both object level rearrangement, fine grained hand object interaction. The objects include everyday items, containers, tools, and packaged goods.

Robot task executions are collected through human teleoperation under a shared collection protocol, except for the five barehand scenarios, which are recorded from a person performing the task bare-handed.The dataset, as part of this benchmark, contains 1,816 successful execution videos and approximately 25 hours of real manipulation data. Repeated executions introduce variation in object configurations, motion trajectories, and interaction patterns within each task in a consistent manner.

Visual observations are recorded from robot mounted or near egocentric cameras. Multi view recordings provide complementary head and wrist views, capturing workspace context and local manipulation details, respectively. Single view recordings are also retained. When multiple camera streams are available, they are synchronized within each execution. All recorded episodes are manually inspected after collection before inclusion in the benchmark and associated with task level metadata.

Each execution is annotated with task-relevant action and state intervals. We represent the annotations as $\mathcal{A} = \{(s_i, e_i, z_i)\}_{i=1}^{K}$, where s_i and e_i denote the start and end times of an interval, and z_i describes the action or task state. Examples include approaching an object, establishing a grasp, moving an object toward a target, completing placement, and reaching a task-specific goal state. We use temporal intervals rather than timestamps to preserve the duration of actions and persistent task states. These annotations provide the reference semantics and temporal boundaries used to sample evaluation points and construct task-specific questions and answers. The benchmark contains 34,713 evaluation instances spanning seven evaluation tasks across 39 manipulation scenarios.

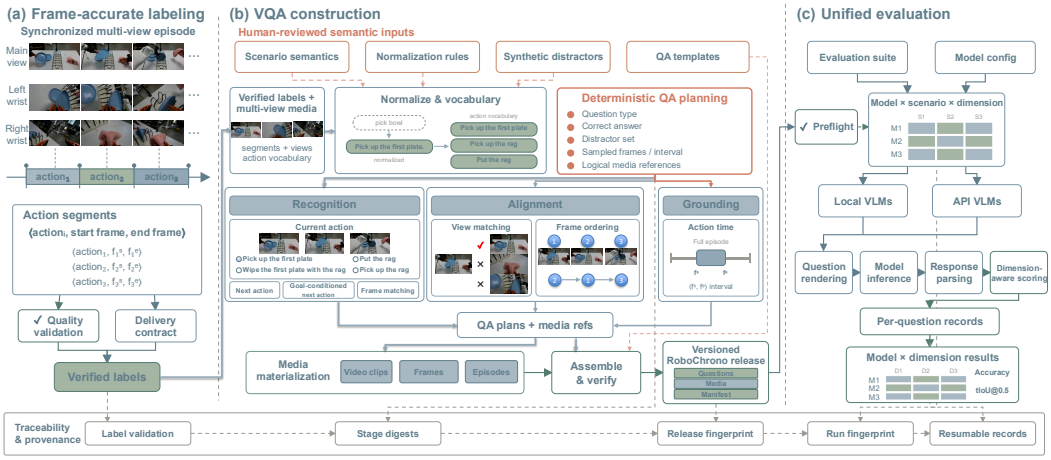

Figure 3: Overview of the RoboChrono construction and evaluation pipeline.

3.3 Evaluation Protocol

We evaluate all models under a causal observation setting that reflects the information available to an embodied agent during online interaction. For an evaluation instance at time t, the model can only access visual observations available at or before t. Future observations are never included in the model input. For next action prediction, later annotations may be used during benchmark construction to determine the reference answer, but the corresponding future visual evidence remains hidden at inference time. Matching, ordering, and temporal localization instances are similarly constructed using only observations available within the permitted observation range.

The seven tasks instantiate this causal protocol with different forms of visual evidence and reference targets. Table 2 summarizes the model inputs and target outputs for each task.

Table 2: Task inputs and reference targets in RoboChrono.

Task	Model input	Reference target
Current action	Video prefix $\mathcal{O}_{\leq t}$ and answer choices	Action or state at time t
Next action	Video prefix $\mathcal{O}_{\leq t}$ and answer choices	Next action executed after t
Goal-conditioned next action	Video prefix $\mathcal{O}_{\leq t}$, task goal, and answer choices	Next executed action under the goal
Frame matching	Query state or frame and candidate frames from $\mathcal{O}_{\leq t}$	Frame matching the queried state
View matching	Query image and synchronized candidates from other views	Image from another view at the same time step
Frame ordering	Candidate frames from $\mathcal{O}_{\leq t}$	Correct temporal order
Action time localization	Video observation, queried action/state, and video duration	Interval $[s_i, e_i]$ in seconds

View matching is generated only when synchronized multi-view recordings are available; single-view recordings are excluded from this task rather than counted as incorrect. When an action label appears multiple times in one execution, the queried instance is disambiguated by its temporal position and annotation index, so that each question refers to a unique annotated interval.

All models are evaluated in a strictly zero-shot setting without benchmark-specific supervised fine-tuning or adaptation. We use the same task definitions, answer spaces, and scoring rules across models. Prompt templates are kept consistent whenever permitted by the corresponding inference interfaces. Closed-source models are evaluated through their officially available interfaces, while open-source models are evaluated locally using their publicly released checkpoints.

For the six multiple-choice tasks, performance is measured using accuracy. Given N valid evaluation instances, the accuracy is computed as

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i = y_i^*\right], \tag{1}$$

where $\hat{y}_i$ is the model prediction and y_i^* is the annotated reference answer. Task–recording combinations for which a question is not applicable, such as view matching for single-view recordings, are excluded from the corresponding evaluation set rather than treated as incorrect predictions.

For action time localization, the model predicts a temporal interval $\hat{I} = [\hat{s}, \hat{e}]$ corresponding to the queried action or task state. The localization prompt specifies the total duration of the video and requires the predicted interval to be expressed in seconds. We measure the overlap between the predicted interval $\hat{I}$ and the annotated reference interval I^* using temporal Intersection over Union:

$$\mathrm{tIoU}(\hat{I}, I^*) = \frac{|\hat{I} \cap I^*|}{|\hat{I} \cup I^*|}, \tag{2}$$

where $|\cdot|$ denotes temporal duration. Since each query requires a single predicted interval, we report Recall@1 at a tIoU threshold of $\tau = 0.5$:

$$R@1(\tau) = \frac{1}{N_{\mathrm{loc}}} \sum_{i=1}^{N_{\mathrm{loc}}} \mathbf{1}\left[\mathrm{tIoU}(\hat{I}_i, I_i^*) \geq \tau\right], \tag{3}$$

where N_{loc} denotes the number of valid localization queries. We refer to this metric as tIoU@0.5 throughout the paper. Finally, we distinguish invalid evaluation runs from valid but low-scoring predictions. A model–task run with a response parse failure rate of 50% or higher is treated as an unsuccessful run and is excluded from reported comparisons. Valid runs are retained regardless of their resulting score, so that poor task performance is not conflated with response parsing failure.

4 Experiments and Results

4.1 Experimental Setup

We evaluate 18 multimodal models on RoboChrono under the zero-shot protocol described in Section 3.3. All models use the same task definitions and scoring rules, without benchmark-specific supervised fine-tuning. We report accuracy for the six multiple-choice tasks and tIoU for Action Time Localization. The multiple-choice aggregate is computed over the applicable questions and does not include Action Time Localization, which is scored separately by tIoU.

4.2 Main Results

Table 3 provides the complete benchmark results. GPT-6-Astra achieves the strongest overall performance by a large margin, reaching 80.5% Choice Avg., compared with 62.8% for RynnBrain1.1-122B-A10B, the strongest open-weight model. Astra leads all seven evaluation dimensions, indicating strong and relatively balanced performance across recognition, prediction, alignment, and temporal grounding. Its strongest result is Frame Matching at 98.3%, showing that it can reliably identify the observation corresponding to the queried visual context. It also performs strongly on Current Action Recognition (89.4%) and Goal-Conditioned Next Action Prediction (83.6%), suggesting effective use of both visual evidence and explicit task goals, though Frame Ordering remains weaker.

However, Astra does not saturate the benchmark. Its performance drops to 76.4% on Next Action Prediction, 68.3% on Frame Ordering, 67.5% on View Matching, and 72.1% on Action Time Localization. These dimensions require reasoning beyond identifying visible content: the model must infer what happens next, recover temporal order, associate observations across viewpoints, or

Table 3: **Results on RoboChrono.** Performance across seven evaluation dimensions grouped into recognition, alignment, and grounding. Group scores are averages over their corresponding dimensions, and Choice Avg. is the mean accuracy over the six multiple-choice tasks.

Model	Size	Recognition					Alignment			Grounding	Overall
		Cur. Act.	Next Act.	Next +Goal	Frame Match	*Group*	View Match	Frame Order	*Group*	Action Time	Choice Avg.
chance floor	–	*0.250*	*0.250*	*0.250*	*0.250*	*0.250*	*0.250*	*0.250*	*0.250*	*0.130*	*0.250*
GPT-6-Astra	–	**0.894**	**0.764**	**0.836**	**0.983**	**0.873**	**0.675**	**0.683**	**0.679**	**0.721**	**0.805**
RynnBrain1.1-122B-A10B	122B	0.752	0.511	0.588	0.954	0.708	0.536	0.329	0.442	0.512	0.628
Doubao-Seed-2.0-Lite	–	0.754	0.538	0.631	0.921	0.717	0.451	0.359	0.409	0.577	0.624
RynnBrain1.1-9B	9B	0.730	0.550	0.599	0.849	0.687	0.444	0.328	0.392	0.405	0.598
Qwen3-VL-235B-A22B	–	0.697	0.471	0.561	0.797	0.637	0.392	0.304	0.352	0.445	0.551
Qwen3-VL-32B	32B	0.709	0.509	0.566	0.720	0.630	0.307	0.329	0.317	0.469	0.536
Qwen3-VL-30B-A3B	30B-A3B	0.684	0.426	0.495	0.853	0.621	0.313	0.290	0.302	0.418	0.525
RynnBrain1.1-2B	2B	0.618	0.390	0.464	0.848	0.587	0.413	0.264	0.346	0.151	0.514
Qwen3-VL-8B	8B	0.653	0.459	0.501	0.690	0.580	0.324	0.284	0.306	0.385	0.498
Qwen3-VL-8B-Thinking	8B	0.608	0.453	0.508	0.715	0.575	0.315	0.277	0.298	0.335	0.492
MiMo-VL-7B-RL	7B	0.570	0.428	0.464	0.778	0.565	0.283	0.294	0.288	0.103	0.482
RynnBrain-2B	2B	0.574	0.377	0.438	0.656	0.516	0.344	0.254	0.303	0.188	0.452
Cosmos-Reason2-2B	2B	0.578	0.435	0.458	0.465	0.486	0.304	0.271	0.289	0.383	0.427
Gemini-3.6-Flash	–	0.284	0.308	0.347	0.245	0.295	0.608	0.496	0.558	0.187	0.374
Qwen3-VL-2B	2B	0.509	0.323	0.335	0.396	0.394	0.289	0.270	0.281	0.223	0.360
SenseNova-SI-1.1-2B	2B	0.409	0.349	0.415	0.398	0.393	0.266	0.286	0.275	0.001	0.358
Cosmos3-Edge-2B	2B	0.503	0.401	0.421	0.237	0.390	0.222	0.204	0.214	0.253	0.339
InternVL3-2B	2B	0.433	0.340	0.423	0.260	0.364	0.258	0.262	0.260	0.023	0.332

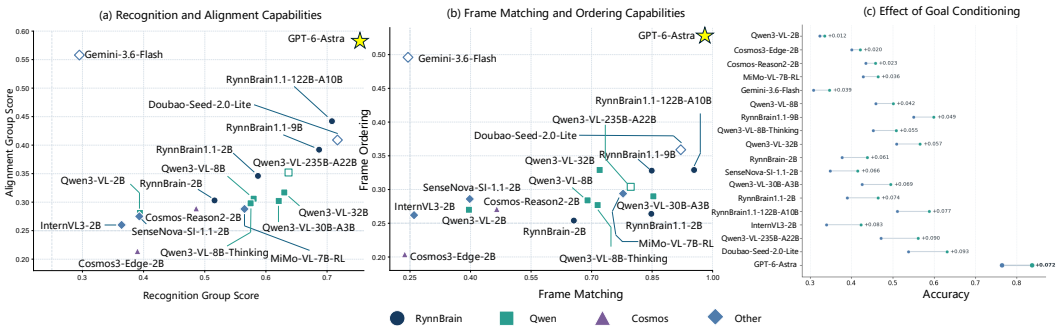

Figure 4: Diagnostic analysis of model capabilities on RoboChrono.

localize an action within the interaction timeline. Thus, even the strongest model retains substantial room for improvement on temporal progress understanding in real-world robot manipulation.

The remaining models exhibit much more uneven capability profiles. RynnBrain1.1-122B-A10B, for example, achieves 95.4% on Frame Matching but only 32.9% on Frame Ordering. This indicates that strong visual correspondence does not necessarily translate into understanding when events occur or in which order they unfold. Doubao-Seed-2.0-Lite shows a similar gap: it performs well on Current Action Recognition (75.4%) and Goal-Conditioned Next Action Prediction (63.1%), but reaches only 35.9% on Frame Ordering and 45.1% on View Matching. RynnBrain1.1-9B is comparatively strong on future prediction among the non-Astra models, obtaining 55.0% on Next Action Prediction, but its Frame Ordering accuracy remains only 32.8%. Together, these results suggest that predicting a plausible next action and recovering the temporal structure of an observed manipulation remain distinct challenges. Gemini-3.6-Flash exhibits a notably different profile. Its recognition performance is weak, with only 24.5% on Frame Matching and a Recognition Group score of 29.5%, yet it achieves the strongest non-Astra results on View Matching (60.8%) and Frame Ordering (49.6%). This contrast is visible in Figure 4(a,b) and shows that the benchmark separates recognition-oriented capability from alignment and temporal order reasoning rather than measuring a single monolithic notion of visual understanding, which aggregate accuracy alone would largely conceal.

Table 4: **Task-matched comparison across robot collection platforms.** We restrict GIM and Tianji gripper recordings to the 12 manipulation tasks available on both platforms and report question-weighted accuracy to isolate platform-specific effects.

Evaluation dimension	GIM	Tianji	Δ	Tianji > GIM (models)
Current Action	54.1	63.7	-9.6	17 / 18
Next Action	39.9	45.0	-5.1	14 / 18
Goal-conditioned Next Action	43.9	48.1	-4.2	14 / 18
Frame Matching	57.2	66.7	-9.6	17 / 18
View Matching	32.5	35.5	-3.0	15 / 18
Frame Ordering	27.1	30.0	-2.9	17 / 18
Choice Avg.	**43.1**	**48.8**	**-5.8**	**18 / 18**

At the lower end of the table, InternVL3-2B, Cosmos3-Edge-2B, SenseNova-SI-1.1-2B, and Qwen3-VL-2B obtain Choice Avg. scores between 33.2% and 36.0%. Their weaknesses are broad rather than task-specific. For example, InternVL3-2B and Cosmos3-Edge-2B remain close to the chance floor on several matching and ordering dimensions, while SenseNova-SI-1.1-2B achieves an Action Time score of only 0.001. These models therefore struggle not only with future prediction but also with basic temporal grounding and cross-observation reasoning.

Figure 4(c) further shows that explicit task goals consistently improve next-action prediction. Astra improves from 76.4% to 83.6%, while all other models also gain from goal conditioning. This indicates that knowing the intended task helps constrain plausible future actions. Nevertheless, the persistent gap between Next Action and the stronger recognition tasks shows that goal information alone is insufficient: models must still connect the current visual state with the temporal progress of the manipulation. Overall, the benchmark reveals a clear hierarchy of difficulty. Models are generally strongest at recognizing or matching visible observations, weaker at predicting future actions, and substantially less reliable at temporal ordering, cross-view alignment, and precise action-time grounding. GPT-6-Astra substantially raises the performance ceiling and shows that strong performance across all capability groups is achievable, but its remaining errors on ordering, viewpoint alignment, prediction, and temporal localization also indicate that real-world manipulation progress understanding is not yet solved, leaving substantial headroom for future models.

4.3 Effect of Robot Collection Platform

To separate platform effects from task composition, we compare GIM and Tianji only on the 12 manipulation tasks recorded on both platforms.

As shown in Table 4 and Figure 5(a), model performance is broadly consistent across the two robot collection platforms. Tianji obtains moderately higher scores on all six matched dimensions, with an overall Choice Avg. difference of 5.8 percentage points, but the magnitude of the shift remains relatively small compared with the overall variation across models and tasks. The largest differences occur in Current Action Recognition and Frame Matching ($\sim$9.6 points), while Frame Ordering differs by only 2.9 points. Importantly, the overall capability pattern remains similar across platforms: dimensions that are difficult on GIM remain difficult on Tianji, and the relative strengths and weaknesses of the evaluated models are largely preserved. We therefore view the observed gap as a moderate sensitivity to collection conditions rather than evidence that benchmark performance is dominated by a particular robot platform or its specific camera and gripper hardware.

4.4 Input Ablation

We further examine which input cues support benchmark performance using five open-weight models on a balanced 312-stem subset. Each ablation is evaluated against the Full condition on matched questions. Figure 5(b) reveals markedly different input dependencies across tasks. Removing visual observations causes large drops in Current Action Recognition and Frame Matching (-22.1 and -30.2 points), whereas Next Action Prediction changes by less than one point. Similarly, replacing the observation history with only the current frame reduces Current Action Recognition by 22.5 points but has only a small effect on future action prediction. This suggests that

Table 5: Input ablation on RoboChrono

Input	Current Action	Next Action	Goal-cond. Next	Frame Match	Frame Order	View Match	Action Time
Full	**62.6**	**40.0**	**49.3**	**46.0**	**29.4**	**34.9**	**30.4**
Blind	40.5 (-22.1)	39.3 (-0.7)	41.1 (-8.2)	15.8 (-30.2)	27.5 (-1.9)	22.8 (-12.1)	7.6 (-22.8)
Text Only	–	–	–	27.6 (-18.4)	–	24.8 (-10.1)	–
Current Frame Only	40.0 (-22.5)	37.5 (-2.6)	46.2 (-3.0)	–	–	–	–
Shuffled Frames	53.6 (-8.9)	39.2 (-0.9)	47.6 (-1.7)	56.4 (+10.4)	–	–	12.1 (-18.3)
Identity Re-encode	62.1 (-0.4)	–	–	45.9 (+0.0)	–	–	–

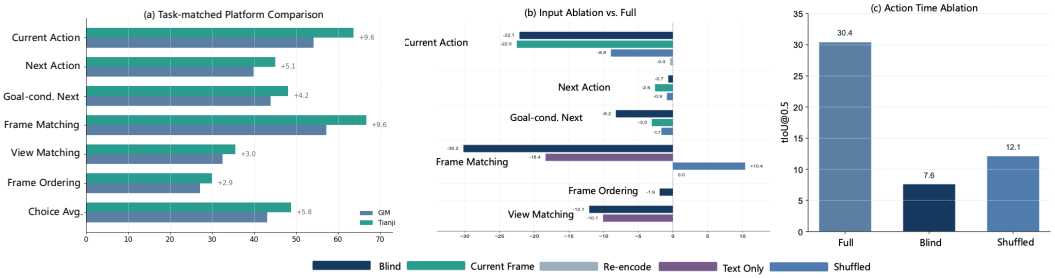

Figure 5: Controlled analysis of collection and input effects.

current action understanding relies strongly on visual history, while next-action prediction can often be supported by task and action priors even when visual evidence is entirely unavailable.

Temporal order produces a complementary effect. Shuffling frames reduces Current Action Recognition by 8.9 points and Action Time Localization from 30.4 to 12.1 (Figure 5(c)), showing that temporal structure is important for grounding task progress. By contrast, Frame Matching increases under shuffling, while the identity re-encoding control produces almost no change. The effect therefore reflects task specific use of temporal order rather than a generic video processing artifact.

Overall, the ablations expose a clear distinction between *recognizing visible evidence* and *understanding task progression*. Visual observations are essential for recognition and matching, temporal context is critical for current state and temporal grounding, while next-action prediction remains comparatively insensitive to the exact visual history, relying more on task priors.

5 Conclusion

We introduced RoboChrono, a real-robot benchmark for evaluating temporal task understanding from progressively revealed visual observations, covering 39 manipulation scenarios and seven complementary evaluation tasks. Our results reveal a persistent gap between recognizing visible content and understanding how a physical task unfolds over time. Even strong multimodal models exhibit substantial weaknesses in temporal ordering, future action prediction, viewpoint alignment, and temporal localization, while controlled ablations further reveal distinct dependencies on observation history, temporal order, and task priors. These findings highlight the importance of evaluating temporal capabilities individually rather than relying solely on aggregate performance. We hope RoboChrono provides a useful testbed for developing multimodal systems that can more reliably reason about task progress during ongoing real-world robot interaction.

AI Use Statement

Generative AI tools were used only for language editing and writing assistance. They were not used for data generation, experimental execution, result analysis, or scientific decision making. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Tianwei Bao, Qineng Wang, Kangrui Wang, Mingkai Deng, Guangyi Liu, Jiayuan Mao, Lawrence Birnbaum, Zhiting Hu, Eric P Xing, Zhaoran Wang, et al. Actioneqa: Action interface for embodied question answering. *Transactions on Machine Learning Research*, 2026.

Kaiyuan Eric Chen, Shuangyu Xie, Zehan Ma, Pannag Sanketi, and Ken Goldberg. Robo2vlm: Improving visual question answering using large-scale robot manipulation data. *Advances in Neural Information Processing Systems*, 38, 2026a.

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

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*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: the dataset. In *European conference on computer vision*, pp. 753–771. Springer, 2018.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026.

Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 6252–6261, 2019.

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.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: A comprehensive benchmark and memory-augmented method. *arXiv preprint arXiv:2501.00584*, 1(2):5, 2024.

Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1724–1734. IEEE, 2025.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, et al. Robobench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. *arXiv preprint arXiv:2510.17801*, 2025.

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.

Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6748–6765, 2021.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780. IEEE, 2025.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

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

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

A APPENDIX

Hardware Platforms. Our benchmark is built upon two real-world robotic manipulation platforms, covering different robot embodiments and workspace configurations. As illustrated in Fig. 6, the first platform is a dual-arm robotic system for whole-workspace manipulation, while the second platform consists of two tabletop robotic manipulators. Both platforms are operated through a VR-based teleoperation interface, where human operators use a head-mounted display and handheld controllers to perform manipulation tasks. These complementary hardware setups provide diverse real-world robot trajectories for benchmark construction and evaluation.



Figure 6: **Hardware platforms used in our benchmark.** We employ two real-world robotic manipulation setups with different embodiments and workspace configurations. Both platforms support VR-based human teleoperation for collecting manipulation trajectories.

Conditional Action Time Evaluation. The Action Time task requires models to answer with temporal intervals in seconds. Although the corrected prompt states the total duration of the queried video, some systems still return intervals as fractions of the video duration. This behavior is a unit error rather than a temporal localization error. We therefore report a conditional Action Time score that evaluates only the answers that are given in seconds.

Let $\mathcal{S}$ denote the set of localization queries for which the model returns a parseable interval in seconds. The conditional score is computed as

$$\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{1} \left[\mathrm{tIoU}(\hat{I}_i, I_i^*) \geq 0.5 \right].$$

The reference interval and scoring threshold are unchanged from the main evaluation. The only difference is that normalized interval answers are excluded from the denominator.

Qualitative Action Time Examples. Figure 7 presents six qualitative examples of Action Time localization from the GIM and Tianji platforms, comparing the predictions of five models with the ground-truth intervals. The examples illustrate both accurate temporal alignment and different failure cases, including overly broad intervals, predictions that do not overlap the reference interval, and timestamps beyond the video duration.



Figure 7: **Qualitative examples of Action Time localization.**

RoboChrono Evaluation Pipeline. Algorithm 1 summarizes the complete RoboChrono evaluation pipeline across different models, scenarios, and evaluation dimensions. For each question, the pipeline prepares the visual input, performs model inference, parses either a multiple-choice answer or a temporal interval, and computes the corresponding accuracy or tIoU@0.5 score. Finally, all question-level results are aggregated into the evaluation report.

Algorithm 1: RoboChrono Evaluation

Input: Models $\mathcal{M}$, scenarios $\mathcal{S}$, dimensions $\mathcal{D}$
Output: Evaluation report $\mathcal{R}$

Validate dataset and model configurations;
Initialize result set $\mathcal{R} \leftarrow \emptyset$;
foreach *model* $m \in \mathcal{M}$ **do**
  Initialize model adapter A_m;
  foreach *scenario* $s \in \mathcal{S}$ **do**
    foreach *dimension* $d \in \mathcal{D}$ **do**
      Load questions $Q_{s,d}$;
      foreach *question* $q \in Q_{s,d}$ **do**
        if *q has already been completed* **then**
          continue;
        $x \leftarrow \mathrm{PrepareInput}(q)$;
        $y \leftarrow A_m.\mathrm{Generate}(x)$;
        if $d =$ `action_time` **then**
          $\hat{t} \leftarrow \mathrm{ParseInterval}(y)$;
          $z \leftarrow \mathbb{I}\big(\mathrm{tIoU}(\hat{t}, t_q) \geq 0.5\big)$;
        else
          $\hat{c} \leftarrow \mathrm{ParseChoice}(y)$;
          $z \leftarrow \mathbb{I}(\hat{c} = c_q)$;
        Save (m, s, d, q, y, z);
      $\mathrm{Score}_{m,s,d} \leftarrow \frac{1}{|Q_{s,d}|} \sum_{q \in Q_{s,d}} z_q$;
Aggregate all scores into $\mathcal{R}$;
return $\mathcal{R}$;