STORM-BENCH: EVALUATING ONLINE VIDEO QA UNDER EVOLVING AND INCOMPLETE EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable online video question answering requires tracking state transitions while selectively abstaining when visual evidence is insufficient. Existing benchmarks focus on static recognition or long-range retrieval, rarely evaluating these coupled capabilities under evolving and incomplete evidence. We present STORM-Bench, comprising 5,736 questions across 630 compact, change-dense episodes spanning five egocentric domains (STORM-Real) and two controlled simulation subsets (STORM-Sim) at 1 FPS. Questions are stratified by a proxy for accumulated change intensity (Low, Medium, High) and query-time answerability (Known, Uncertain). To measure reliability, we introduce STORM-BR, a harmonic metric over joint answer–status correctness that exposes abstention failures masked by aggregate accuracy, alongside STORM-BR-ATTR for uncertainty attribution. Across 14 video LLMs, online accuracy peaks at 60.3% (mean 51.7%), whereas STORM-BR ranges from 5.7% to 35.6% (mean 18.8%), driven by pervasive overconfidence on uncertain queries. STORM-Bench shows that task accuracy masks these gaps in epistemic reliability and state tracking. Evaluation code is available at `https://anonymous.4open.science/r/STORM-Bench/`.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have expanded video question answering (Video-QA) beyond static clip recognition (Li et al., 2024; Fu et al., 2025). Standard offline benchmarks evaluate evidence integration after the entire video has been observed. Real-world interactive scenarios instead demand **online Video-QA**, where questions are associated with specific timestamps and visual input is restricted to the history available by each query (Lin et al., 2024; Yang et al., 2025; Niu et al., 2025). Under evolving and incomplete evidence, access to historical frames does not guarantee query-time answerability: earlier observations can become outdated after dynamic state changes, while the latest visual evidence may remain occluded or unobserved.

Recent benchmarks expand the temporal horizon across extended, hour-scale videos (Mangalam et al., 2023; Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025a). We study a complementary, under-explored challenge: dense scene transitions within compact videos, where frequent relocations, transformations, and replacements force a model to update past evidence and integrate multiple state flips in a short window. STORM-Bench builds such episodes from repeated visits to a shared functional region (Figure 2b). As Figure 3 shows, its episodes combine compact durations (a) with higher sampled visual-change rates (b) and denser QA supervision (c) than the compared suites, making belief updating and principled abstention the primary test.

Dense state changes further require **principled abstention under insufficient evidence**. Existing state-tracking benchmarks examine evolving answers and event boundaries in dynamic scenes (Gao et al., 2021; Xun et al., 2025; Liu et al., 2026; Forte et al., 2026), while selective-abstention suites introduce missing frames, linguistic ambiguity, or physical interventions (Yu et al., 2026; Wu et al., 2026; Pramono et al., 2026; Azad et al., 2026). In dynamic online interaction, however, these challenges co-occur: a model must update its internal state upon observed transitions while withholding answers when critical evidence is occluded or missing. Evaluating either in isolation conceals failures where stale observations and evidence gaps interact.

We introduce **STORM-Bench** to jointly evaluate these capabilities: an online video benchmark of 5,736 questions across 630 compact, change-dense episodes at 1 FPS. STORM-Real combines Cook

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.

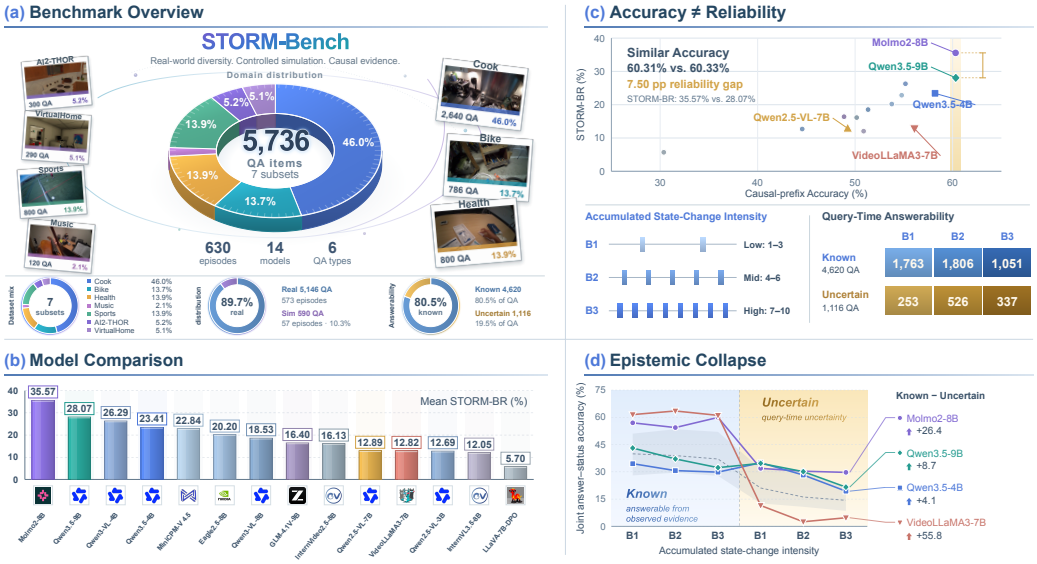

Figure 1: **Overview of STORM-Bench.** (a) Domain mix with Real/Sim and Known/Uncertain splits. (b) STORM-BR ranking of 14 video LLMs. (c) Nearly matched accuracy hides a 7.50-point STORM-BR gap. (d) By change intensity: Known remains stronger for most models; Uncertain often collapses under overconfidence (seven-subset B_1–B_3 average; cf. Table 20, Figure 7).

from HD-EPIC (Perrett et al., 2025) with four Ego-Exo4D domains (Grauman et al., 2024), and STORM-Sim adds THOR (AI2-THOR (Kolve et al., 2017)) and VHome (VirtualHome (Puig et al., 2018)) (Figure 2a). Questions are stratified along two diagnostic axes (Figure 2c): **accumulated change intensity** (Low, Medium, High), counting initiated visit or event intervals prior to the query, and **query-time answerability** (*Known* vs. *Uncertain*), indicating whether observed history supports a decisive answer or mandates abstention.

We assess balanced reliability using STORM-BR, a Laplace-smoothed harmonic mean over joint answer–status correctness that exposes abstention failures masked by aggregate accuracy, complemented by STORM-BR-ATTR for uncertainty-source attribution. Across 14 open-weight video LLMs, online task accuracy peaks at 60.3% (mean 51.7%), whereas STORM-BR ranges from 5.7% to 35.6% (mean 18.8%). Nearly identical accuracy can conceal up to a threefold reliability gap, driven by severe status overconfidence on Uncertain items. Furthermore, models struggle on history aggregation and exhibit lower performance on questions whose evidence is older than five seconds. These results show that standard task accuracy alone does not capture the state-tracking and answerability failures observed under evolving evidence. This work makes three contributions:

- **A change-dense video benchmark.** We build STORM-Bench, concentrating physical transitions and repeated visits into compact episodes (39–67 s) across five real domains and two controlled simulation subsets, achieving higher sampled visual-change and QA density than prior suites.
- **A balanced reliability formulation.** We formalize query-time answerability via decoupled probing and introduce STORM-BR and STORM-BR-ATTR to summarize joint answer–status correctness across the intensity–answerability grid and uncertainty-source attribution.
- **In-depth diagnostics across 14 video LLMs.** We uncover a pervasive overconfidence bottleneck on Uncertain queries across most models, lower joint accuracy on older-evidence queries (>5 s), and a pronounced vulnerability in history aggregation.

2 Preliminaries and Related Work

2.1 Problem Formulation: Evolving Evidence and Answerability

Consider an online video stream $V = \{(\tau_j, v_j)\}_{j\geq 1}$, where visual observation v_j arrives at monotonically increasing timestamp τ_j. The physical environment undergoes continuous state transitions, represented by a latent state trajectory $S(t) \in \mathcal{S}$. At query timestamp t_q, the model receives a

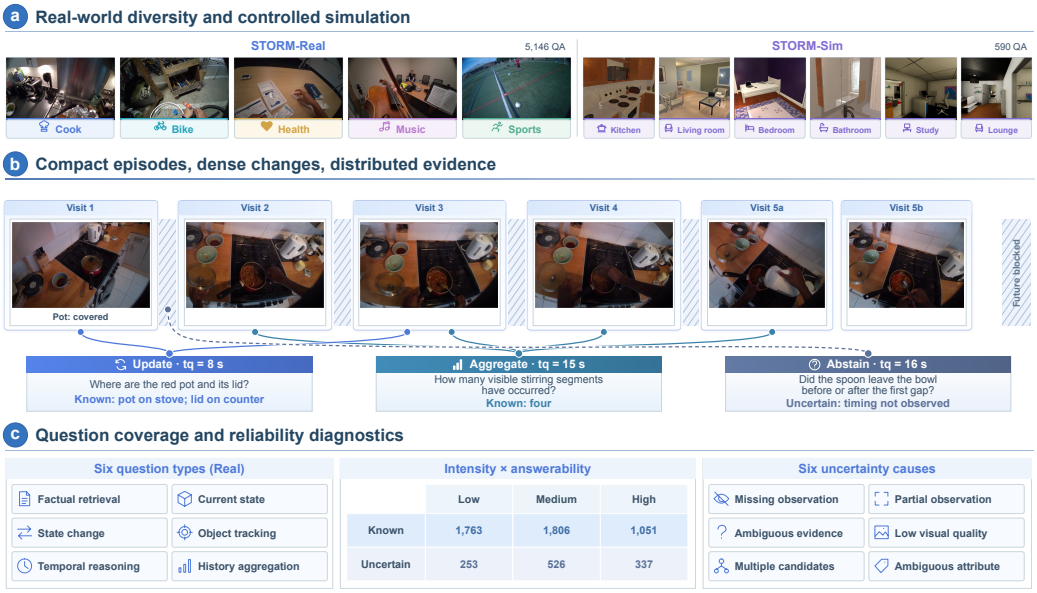

Figure 2: **Overview of STORM-Bench.** (a) Five real egocentric domains and four controlled simulation room categories. (b) Compact multi-visit episodes requiring update, aggregation, and abstention. (c) Six question types stratified by intensity×answerability, with uncertainty causes.

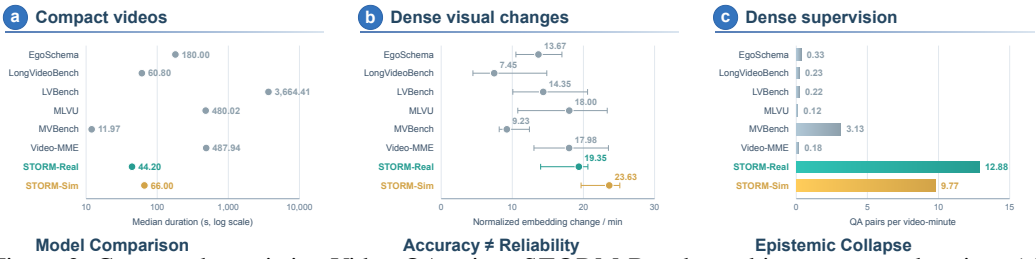

Figure 3: Compared to existing Video-QA suites, STORM-Bench combines compact durations (a), higher sampled visual-change rates (b), and user QA supervision (c). Dots/bars denote medians; error bars in (b) show IQRs. See Appendix A.5.1.

question q along with the visual history observed up to that moment:

$$V_{\leq t_q^o} = \{(\tau_j, v_j) \in V \mid \tau_j \leq t_q\}, \quad (1)$$

while all future observations $V_{>t_q}$ are masked. We use *online* to denote this setting in which only frames up to the query time are visible. The response space is formulated as $\mathcal{Y}_q \cup \{\perp\}$, where $\mathcal{Y}_q$ denotes the set of substantive candidate options and $\perp$ designates that visual evidence is insufficient.

Evolving Evidence and Query-Time Answerability. Unlike static Video-QA with invariant visual evidence, online video understanding features *evolving evidence* shaped by state transitions:

- **State Updates (Known):** When an entity undergoes a transition $S(t_{\text{prev}}) \to S(t_{\text{new}})$ within $V_{\leq t_q}$ ($t_{\text{prev}} < t_{\text{new}} \leq t_q$), earlier frames capturing $S(t_{\text{prev}})$ become *stale evidence*. A query regarding the current state at t_q (or an observed historical state) is *Known* ($y^* \in \mathcal{Y}_q$) when $V_{\leq t_q}$ unambiguously supports the target fact, requiring the model to update its prediction and suppress outdated history.
- **Observation Gaps and Selective Abstention (Uncertain):** A query is *Uncertain* ($y^* = \perp$) if critical transitions occur out of view, suffer from perceptual occlusion, or lack definitive visual support up to t_q. Under such gaps, $V_{\leq t_q}$ does not uniquely determine a substantive answer. Reliable behavior strictly mandates abstaining via $\perp$ rather than overconfidently guessing.

Section 3 instantiates this formulation in compact episodes with a four-choice set $\mathcal{A} = \{A, B, C, D\}$, where Known items comprise four substantive candidates, and Uncertain items include three alongside an explicit abstention option. To remove task-option cues from answerability assessment, a decoupled status probe evaluates answerability without presenting task candidate options.

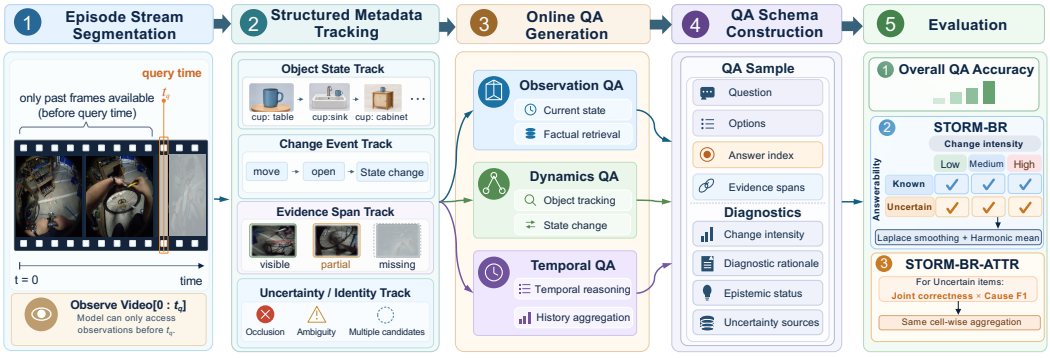

Figure 4: **STORM-Bench diagnostic framework.** Questions posed at query time t_q are answered strictly from the observed history $V_{\leq t_q}$ across six semantic types and two diagnostic axes. Balanced reliability and fine-grained root-cause attribution are scored via STORM-BR and STORM-BR-ATTR.

2.2 Related Work

Long-Context Retrieval vs. Online Video Understanding. A prominent direction in video benchmarking expands the temporal horizon, evaluating long-range needle retrieval and multi-event reasoning across hour-scale videos (Mangalam et al., 2023; Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025a; Fu et al., 2025). In contrast, online Video-QA restricts visual access strictly to the history observed prior to each query (Lin et al., 2024; Yang et al., 2025; Niu et al., 2025; Zhang et al., 2024; Huang et al., 2025). STORM-Bench complements both paradigms by targeting **change-dense temporal reasoning within compact episodes** (39–67 s), concentrating repeated visits, fast relocations, and visual state transitions into short observation windows (Appendix Table 9).

Dynamic State Tracking and Principled Abstention. Existing diagnostic suites evaluate either dynamic entity tracking (Gao et al., 2021; Xun et al., 2025; Liu et al., 2026; Forte et al., 2026; Lei et al., 2026) or selective abstention under missing frames, ambiguity, or physical perturbations (Yu et al., 2026; Wu et al., 2026; Pramono et al., 2026; Azad et al., 2026). In dynamic online interaction, however, these challenges co-occur: a model must update internal states upon observed transitions while withholding predictions when critical evidence is occluded or missing. STORM-Bench unifies these dual demands by cross-stratifying accumulated change intensity with query-time answerability.

Evaluating Video LLMs under Incomplete Observations. Video LLMs handle extended contexts through visual token compression (Wang et al., 2025c; Zhang et al., 2025; LLaVA-VL Team, 2024; Yu et al., 2025) or native temporal embeddings (Bai et al., 2025a;b; Wang et al., 2025b; Chen et al., 2025; Clark et al., 2026). While selective prediction frameworks evaluate coverage–risk trade-offs (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Pramono et al., 2026; Azad et al., 2026), standard accuracy masks catastrophic overconfidence under observation gaps. We evaluate 14 open-weight checkpoints across architectures, using balanced harmonic metrics (STORM-BR, STORM-BR-ATTR) to compare state-related errors and overconfidence under evolving evidence (Appendix C).

3 STORM-Bench

STORM-Bench operationalizes the Video-QA formulation in Section 2.1 into a suite (Figure 4), spanning episode segmentation, metadata annotation, stratification across intensity and answerability, and reliability evaluation via harmonic metrics (STORM-BR, STORM-BR-ATTR).

3.1 Diagnostic Space

While instances span six semantic question types (Figure 4, Table 4), online reliability is critically governed by environmental dynamics and evidence sufficiency. To decouple task semantics from reliability stressors, we stratify all questions across two complementary diagnostic axes.

Accumulated change intensity ($b \in \{B_1, B_2, B_3\}$). To evaluate how dynamic history challenges temporal persistence and belief updating, we assign an ordinal intensity index $c_i \in \{1, \ldots, 10\}$ at query time t_q as a proxy for accumulated scene activity and state evolution. In the real track, c_i

counts initiated revisit intervals to the functional region; in the simulation track, it counts scheduled physical event intervals whose onset boundary is no later than t_q (Appendix A.3). We partition this scale into fixed Low ($B_1 = [1, 3]$), Medium ($B_2 = [4, 6]$), and High ($B_3 = [7, 10]$) bins. Sensitivity audits (Appendix D.4) show greater ranking variation on the simulation subsets than on the larger real-world subsets.

Query-time answerability ($s \in \{\text{Known}, \text{Uncertain}\}$). Following Section 2.1, queries are classified by whether $V_{\leq t_q}$ provides sufficient visual evidence: Known items admit a unique answer in $\mathcal{A}$, whereas Uncertain items lack decisive visual support (due to occlusions or observation gaps) and mandate abstention ($\perp$). Each uncertain query is annotated with one or more of six co-occurring root causes (Appendix A.3): (1) *missing observation*, (2) *partial observation or occlusion*, (3) *ambiguous evidence*, (4) *low visual quality*, (5) *multiple plausible candidates*, or (6) *ambiguous entity attribute*, enabling separate evaluation of uncertainty detection and attribution.

3.2 Decoupled Probing and Reliability Metrics

Each question presents four options $\mathcal{A} = \{A, B, C, D\}$ to predict $\hat{a}_i$. For an Uncertain item, one designated option expresses evidence insufficiency ($\perp$). Relying solely on option selection allows an option-structure heuristic (selecting abstention whenever present, guessing uniformly otherwise) to reach non-trivial accuracy (39.72% expected on Cook; Appendix A.7). To remove task-option cues from status evaluation, STORM-Bench implements a two-stage decoupled probing protocol immediately following primary inference:

- **Stage 1: Decoupled evidence sufficiency probe.** Holding history $V_{\leq t_q}$ and question fixed, the model receives a four-way probe devoid of task options $\mathcal{A}$: (A) directly observed; (B) inferred via temporal reasoning; (C) observations missing; or (D) observations ambiguous. Options A and B map to predicted status $\hat{s}_i = \text{Known}$, whereas C and D map to $\hat{s}_i = \text{Uncertain}$.
- **Stage 2: Fine-grained attribution probe.** When $\hat{s}_i = \text{Uncertain}$, a multi-select probe identifies predicted root causes $\hat{C}_i \subseteq \{1, \ldots, 6\}$ among the six predefined sources; otherwise $\hat{C}_i = \varnothing$.

Joint task-and-status correctness (J_i). To penalize lucky guesses and blind abstentions, credit is awarded strictly when both the primary answer and probed status match ground truth:

$$J_i = \mathbb{1}[\hat{a}_i = a_i \wedge \hat{s}_i = s_i]. \tag{2}$$

Models must simultaneously select the substantive answer and confirm evidence for Known queries, or select the abstention option and independently verify incompleteness for Uncertain queries.

Balanced reliability score (STORM-BR). Crossing three change-intensity bins with two answerability states yields a 3×2 diagnostic grid. Let $I_{b,s} = \{i : c_i \in b, s_i = s\}$ index instances in cell (b, s). Online cell frequencies are inherently skewed (e.g., $n = 8$ in Bike B_3-Uncertain). Sample-weighted averaging allows frequent cells to mask failures on less frequent conditions, while unregularized harmonic means collapse to zero if any sparse cell fails.

Following harmonic averaging in generalized zero-shot learning (Xian et al., 2017), **STORM-BR** applies Laplace add-1 smoothing (a uniform $\text{Beta}(1, 1)$ prior) to each occupied cell and computes their harmonic mean across all valid conditions $\mathcal{C} = \{(b, s) : |I_{b,s}| > 0\}$:

$$\widetilde{J}_{b,s} = \frac{\sum_{i \in I_{b,s}} J_i + 1}{|I_{b,s}| + 2}, \qquad \text{STORM-BR} = \frac{|\mathcal{C}|}{\sum_{(b,s) \in \mathcal{C}} \widetilde{J}_{b,s}^{-1}}. \tag{3}$$

Cook, Bike, Health, Sports, THOR, and VHome populate all six cells ($|\mathcal{C}| = 6$), while Music lacks high-intensity events and is evaluated over its four occupied cells ($|\mathcal{C}| = 4$).

Diagnostic attribution score (STORM-BR-ATTR). **STORM-BR-ATTR** evaluates whether models accurately discern *why* evidence is insufficient by weighting joint correctness on uncertain items by the multi-label attribution F1 score between predicted causes $\hat{C}_i$ and ground-truth causes C_i:

$$F_i^{\text{cause}} = \frac{2|\hat{C}_i \cap C_i|}{|\hat{C}_i| + |C_i|}, \qquad A_i = \begin{cases} J_i \cdot F_i^{\text{cause}}, & s_i = \text{Uncertain}, \\ J_i, & s_i = \text{Known}. \end{cases} \tag{4}$$

The cell attribution score is $\bar{A}_{b,s} = |I_{b,s}|^{-1} \sum_{i \in I_{b,s}} A_i$. We compute STORM-BR-ATTR as the harmonic mean of these unsmoothed scores over occupied cells, with a value of zero when any

occupied cell has $\bar{A}_{b,s} = 0$. This score measures joint answer–status correctness weighted by attribution F1. Because the cause probe is gated on a predicted `Uncertain` status, the reported cause F1 is an end-to-end uncertainty-source measure that includes status-gating errors.

Status overconfidence rate (OC). To quantify a model's propensity to hallucinate certainty on unanswerable inputs, we measure the status overconfidence rate:

$$\mathrm{OC} = \frac{\sum_{i:s_i=\mathrm{Uncertain}} \mathbb{1}[\hat{s}_i = \mathrm{Known}]}{|\{i : s_i = \mathrm{Uncertain}\}|}. \tag{5}$$

Confidence intervals are estimated via 1,000 episode-clustered bootstrap iterations.

3.3 Episode Construction and Validation

STORM-Bench comprises 5,736 questions across 630 compact episodes spanning real egocentric videos (STORM-Real) and controlled simulations (STORM-Sim). Both tracks share the online protocol and diagnostic bins; question builders are detailed in Tables 4 and 3.

Real-world egocentric episodes (STORM-Real). The real-world track combines Cook from HD-EPIC (Perrett et al., 2025) with Bike Repair, Health, Music, and Sports from Ego-Exo4D (Grauman et al., 2024). To evaluate temporal persistence within practical context boundaries, we compile compact revisit streams from temporally separated observations of shared functional regions (Figure 9, Appendix A.1). Source videos sampled at 1 FPS are segmented by a VLM into region-activity timelines tracking functional zones, active manipulations, and visible state transitions (e.g., relocation, transformation, replacement). An LLM planner clusters segments by functional region to propose multi-visit sequences, assembling 3–10 chronologically ordered clips (sped up by $1.5\times$ and separated by 0.2-second transition buffers) into compact episodes targeting 45 seconds (mean durations 39.00–42.86 s; Appendix A.1.2). Finally, evidence-grounded QA pairs are constructed across the six semantic types (Table 4) with exact evidence visibility spans $W_i \subseteq [0, t_q]$ and diagnostic labels.

Controlled physical simulations (STORM-Sim). STORM-Sim combines THOR (300 QA / 28 episodes) and VHome (290 QA / 29 episodes), built with AI2-THOR (Kolve et al., 2017) and VirtualHome (Puig et al., 2018). Both cover Kitchen, Living Room, Bedroom, and Bathroom. The four stages in Figure 10 separate physical events from what the observer can establish:

- **Controlled rollout.** Seeded camera routes create observation gaps around ten scheduled events. THOR uses disappearance/reappearance interventions; VHome uses native open/close, switch, and pickup/placement actions by separate operating agents.
- **Simulator-grounded event extraction.** Execution traces and rendered masks link events to visible before/after observations. For VHome, visibility transitions require two consecutive 1 FPS samples; RGB endpoint review supports physical-state questions.
- **Causal evidence selection.** Query times and evidence spans obey $W_i \subseteq [0, t_q]$, with intensity counting initiated event intervals (Eq. 6). Visible absence can resolve a visibility question while leaving a hidden physical state or action Uncertain.
- **Temporally grounded QA.** Deterministic builders instantiate six question families from VLM templates and attach options, evidence, and epistemic labels. VHome includes an abstention option on every item to avoid status cues. Subset-specific rules and validation appear in Appendix A.2.

Auditing and Controls Against Non-Visual Shortcuts. To assess visual dependence and reduce non-visual shortcuts, we institute multi-tiered controls (Appendix A.7): (1) *Linguistic Debiasing*: Substantive options are edited for comparable length, syntactic structure, and lexical overlap, while the designated abstention choice retains standardized phrasing (Appendix A.7). Answer positions are balanced across candidates, with deterministic ordering to reduce positional bias. (2) *Decoupled Probing*: Omitting task options from the status probe prevents models from exploiting lexical cues between options and uncertainty labels. (3) *Empirical Shortcut Controls*: We benchmark blind text-only and shuffled frame controls (Section 4, Table 26). Text-only STORM-BR yields a median of 1.22% (up to 29.75% for MiniCPM-V 4.5), showing model-dependent performance without vision.

4 Experiments

We organise the evaluation around four questions: whole-bench performance across scales (**RQ1**, §4.2), metric and protocol effects on reliability (**RQ2**, §4.3), answerability and overconfidence root causes (**RQ3**, §4.4), and temporal dynamics with input controls (**RQ4**, §4.5).

Table 1: **Main evaluation on STORM-Bench (RQ1).** Online balanced reliability (*BR*, Laplace-smoothed STORM-BR) per domain with 95% bootstrap intervals, overall average BR (*Avg. BR*), and accuracy (*Avg. Acc*). Both are unweighted means across five real domains, THOR, and VHome.

Model	STORM-BR by domain							Avg. BR	Avg. Acc
	Cook	Bike	Health	Music	Sports	THOR	VHome		
Molmo2-8B (2026)	**34.45** [27.4, 39.3]	**33.59** [22.7, 40.4]	31.80 [18.4, 38.8]	**43.29** [24.2, 55.4]	**31.34** [21.8, 36.8]	**34.15** [21.6, 42.4]	**40.35** [31.8, 44.9]	**35.57**	60.31
Qwen3.5-9B (2026)	33.29 [27.7, 37.8]	22.22 [12.2, 25.7]	33.23 [25.0, 38.7]	42.70 [25.1, 52.4]	24.06 [18.5, 28.2]	25.89 [12.6, 33.2]	15.10 [9.5, 18.1]	28.07	**60.33**
Qwen3-VL-4B (2025b)	26.77 [21.9, 30.4]	31.49 [23.1, 35.3]	29.71 [21.3, 34.8]	36.92 [20.9, 44.6]	20.89 [13.4, 25.2]	19.27 [10.8, 27.7]	18.96 [11.0, 24.9]	26.29	55.16
Qwen3.5-4B (2026)	34.04 [28.8, 38.5]	8.72 [6.3, 12.3]	**33.95** [23.8, 39.9]	36.97 [22.9, 46.0]	18.09 [11.0, 22.2]	15.42 [8.3, 20.0]	16.65 [8.1, 20.7]	23.41	58.20
MiniCPM-V 4.5 (2025)	22.17 [16.5, 26.0]	31.97 [22.8, 37.1]	18.71 [8.5, 25.9]	27.89 [14.9, 38.8]	17.16 [8.5, 22.3]	21.70 [11.5, 26.2]	20.33 [10.9, 27.5]	22.84	54.78
Eagle2.5-8B (2025)	23.63 [18.3, 27.5]	28.86 [20.1, 34.8]	15.74 [8.0, 20.8]	25.85 [13.9, 34.2]	12.41 [6.7, 16.5]	15.44 [7.6, 21.4]	19.45 [11.3, 22.8]	20.20	53.76
Qwen3-VL-8B (2025b)	15.21 [10.5, 18.8]	19.00 [10.3, 24.0]	21.28 [12.4, 26.3]	27.92 [14.0, 39.1]	16.05 [8.6, 20.5]	11.23 [8.0, 12.9]	19.04 [10.2, 25.4]	18.53	51.33
GLM-4.1V-9B-Thinking (2025)	21.40 [19.1, 23.3]	13.35 [8.0, 16.3]	21.69 [17.8, 24.2]	22.21 [11.2, 29.8]	10.76 [7.1, 12.9]	11.34 [8.8, 13.0]	14.03 [5.8, 19.9]	16.40	48.86
InternVideo2.5-8B (2025c)	13.39 [9.3, 16.2]	20.72 [13.9, 24.3]	12.27 [6.4, 17.1]	26.10 [14.4, 33.2]	14.37 [7.2, 18.2]	10.64 [7.7, 12.8]	15.45 [7.7, 19.9]	16.13	50.16
Qwen2.5-VL-7B (2025a)	31.60 [26.8, 35.5]	3.77 [1.5, 5.1]	7.78 [4.0, 9.9]	10.69 [4.8, 15.8]	4.73 [2.7, 6.1]	10.17 [7.4, 11.6]	21.50 [14.0, 26.5]	12.89	49.24
VideoLLaMA3-7B (2025)	11.10 [6.5, 14.6]	15.71 [6.2, 21.5]	9.66 [4.3, 14.5]	18.72 [12.9, 26.0]	15.93 [8.0, 20.6]	9.32 [7.6, 10.7]	9.31 [8.8, 9.9]	12.82	56.09
Qwen2.5-VL-3B (2025a)	19.22 [13.5, 23.1]	12.69 [9.4, 14.4]	13.48 [9.7, 15.9]	10.30 [4.6, 14.6]	9.62 [6.2, 11.5]	10.30 [7.7, 11.9]	13.25 [8.8, 15.6]	12.69	44.57
InternVL3.5-8B (2025b)	8.97 [5.1, 11.7]	7.41 [3.8, 11.6]	17.14 [8.3, 22.2]	16.05 [10.4, 22.9]	8.29 [4.5, 11.0]	14.00 [8.4, 17.2]	12.50 [8.3, 17.4]	12.05	50.87
LLaVA-NeXT-Video-7B-DPO (2024)	1.13 [1.1, 1.2]	3.72 [3.5, 4.0]	3.82 [3.2, 4.3]	11.48 [9.5, 14.0]	3.43 [3.1, 3.8]	8.23 [6.5, 9.4]	8.10 [7.4, 8.6]	5.70	30.37

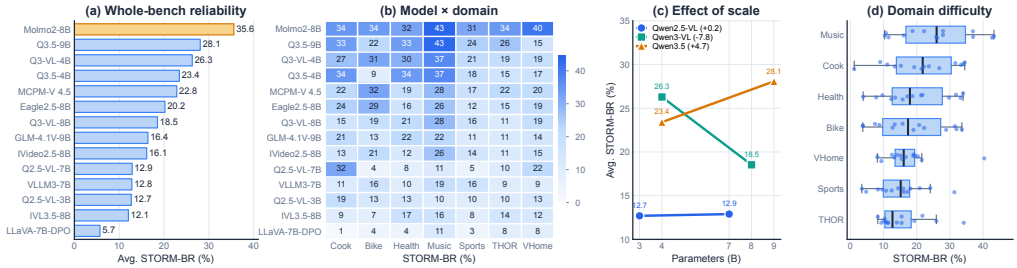

Figure 5: **Whole-bench evaluation across five real domains and two simulation subsets.** (a) Average STORM-BR across seven evaluation subsets. (b) Per-domain STORM-BR scores. (c) Parameter scaling across matched model families. (d) Domain difficulty distribution across all 14 models.

4.1 Experimental Setup

Evaluation tracks. STORM-Bench (Section 3.3) benchmarks 5,736 questions across seven subsets: five real tracks from STORM-Real (Cook: 2,640, Bike: 786, Health: 800, Music: 120, Sports: 800; 6.66h) and simulations in THOR (300, 0.47h) and VHome (290, 0.54h), with domains evaluated independently under the causal-prefix protocol (Appendix A.5, Tables 5, 6).

Evaluation protocol and metrics. We evaluate models under the default *online* protocol ($V_{\leq t_q}$) and an unconstrained *offline* control (V). Standard models decode greedily up to 8 tokens; GLM uses a 512-token thinking budget with separate answer completion. Frames are resized to a 200,704-pixel budget, retaining latest frames under context limits. Decoupled probes execute sequentially after primary inference without task option leakage. Primary evaluation relies on STORM-BR (Eq. 3), STORM-BR-ATTR, and overconfidence (OC), while Kendall's τ, paired exact sign tests, and 1,000 episode-clustered bootstrap intervals quantify statistical concordance and uncertainty.

Evaluated models. We benchmark 14 open-weight checkpoints across 11 architectures: Qwen2.5-VL (3B/7B) (Bai et al., 2025a), Qwen3-VL (4B/8B) (Bai et al., 2025b), Qwen3.5 (4B/9B) (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025b), Molmo2-8B (Clark et al., 2026), MiniCPM-V 4.5 (Yu et al., 2025), InternVideo2.5-8B (Wang et al., 2025c), Eagle2.5-8B (Chen et al., 2025), VideoLLaMA3-7B (Zhang et al., 2025), LLaVA-NeXT-Video-7B-DPO (LLaVA-VL Team, 2024), and GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025). All evaluations are zero-shot end-to-end under native input limits; Appendix B.3 details environments, prompts, and probes.

4.2 RQ1: Whole-Bench Evaluation

A crossed metric separates an otherwise compressed leaderboard. Across seven subsets (Table 1, Figure 5a), Molmo2-8B leads average STORM-BR (35.57%; accuracy 60.31%), followed by Qwen3.5-9B (28.07% / 60.33%); LLaVA-NeXT-Video-7B ranks last (5.70% / 30.37%). Domain difficulty varies: median STORM-BR peaks on Music (25.97%) and Cook (21.78%), bottoming on THOR

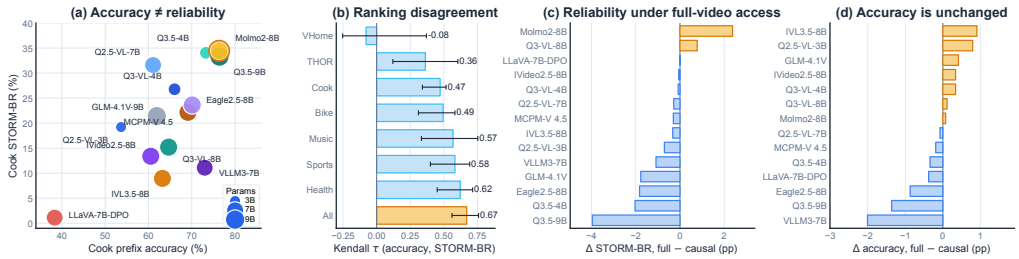

Figure 6: **Diagnostic protocol and metric analysis (Cook domain except panel b).** (a) Online accuracy vs. STORM-BR (marker size $\propto$ scale). (b) Per-domain Kendall's τ (95% bootstrap CIs; highlighted bar pools domains). (c,d) Shift in STORM-BR and accuracy under full-video access.

(12.67%; VHome: 16.05%). The harmonic formulation exposes domain specialization obscured by aggregate accuracy; e.g., Qwen3.5-4B excels on Health (33.95%) yet falls to 8.72% on Bike Repair.

Parameter scaling does not consistently improve balanced reliability. Among the three matched model families (Figure 5c), Qwen3.5-9B gains 2.13 accuracy points and 4.67 STORM-BR points over Qwen3.5-4B. Qwen2.5-VL-7B gains 4.66 accuracy points over its 3B counterpart but only 0.20 STORM-BR points. Qwen3-VL-8B degrades on both accuracy (55.16% to 51.33%) and STORM-BR (26.29% to 18.53%) relative to the 4B checkpoint. At a matched 8B scale, Molmo2-8B achieves 35.57% average STORM-BR, nearly tripling InternVL3.5-8B (12.05%). Parameter scaling alone therefore fails to guarantee improved reliability when reasoning over dynamic physical observations.

Cross-domain ranking stability. Across 10 STORM-Real domain pairs, rank concordance is $\tau_b = 0.72$ (Known) and 0.77 (Uncertain) for joint accuracy vs. 0.55 for aggregate STORM-BR (Appendix Tables 16, 17); five-domain Kendall's W is 0.87, 0.92, and 0.75. Individual answer and abstention capabilities transfer more consistently than their joint harmonic balance. When transferring from Cook to THOR, Known accuracy drops across all 14 models (median -26.23), while Uncertain accuracy rises for 9 models (median $+4.25$) and overconfidence declines for 10. VHome also reduces median Known joint accuracy (-39.74), while its median Uncertain shift is $+3.34$, describing distinct prediction profiles across real and simulated environments (Table 18).

4.3 RQ2: Protocol and Metric Effects

Similar online accuracy hides a threefold reliability gap. On Cook, STORM-BR separates models with near-identical accuracy (Figure 6a): Qwen3.5-4B and VideoLLaMA3-7B reach near-identical accuracy (73.26% vs. 72.99%), yet STORM-BR differs threefold (34.04% vs. 11.10%) under joint answer–status evaluation. Kendall's τ between accuracy and STORM-BR is 0.47 on Cook (inverting 24 of 91 pairs) and varies: Bike (0.49), Health (0.62), Music (0.57), Sports (0.58), THOR (0.36), and VHome (-0.08; Figure 6b). Standard accuracy is insufficient for online reliability.

Full-video access under fixed query-time labels. In the offline control, models receive the full video while evaluating questions under original query-time ground truth. For the original real domains and THOR, evaluation retains the question text without adding a query-boundary marker (Appendix B.3); unsegmented future frames can both provide extraneous context and shift temporal reference for relative queries. Under this protocol stress test, Cook accuracy changes by ≤ 2.01 points (Figure 6d), but STORM-BR declines for 11 models (median -0.31; Qwen3.5-9B drops 3.96 points; Figure 6c, Table 15), showing sensitivity to the protocol change under fixed query-time labels.

4.4 RQ3: Answerability and Overconfidence Diagnosis

Answerability creates a wider performance divide than change intensity. Across all seven subsets, joint accuracy on Known queries exceeds Uncertain queries for 12 of 14 models (median gap 21.82 points, domain-macro Table 20; Figure 1d), with Qwen2.5-VL-7B and GLM reversing this trend. On Cook (Figure 7), median joint accuracy collapses from 68.71% (Known) to 12.74% (Uncertain), while Known accuracy declines by 8.18 points from Low (B_1) to High (B_3) intensity. Uncertain items are the primary bottleneck for 13 of 14 models, driven by overconfidence (median OC $= 52.22\%$). VideoLLaMA3-7B reaches 76.26% on high-intensity Known but only 4.25% on Uncertain items (86.87% OC); GLM-4.1V shows extreme conservatism (OC $= 5.79\%$, suppressing low-intensity Known accuracy to 24.50%; Figure 7d), separating overconfidence from uncertainty.

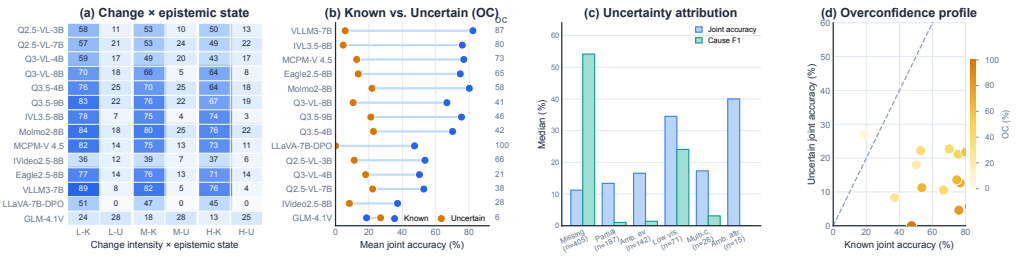

Figure 7: **Answerability diagnostics and error analysis on Cook.** (a) Joint correctness across intensity × answerability cells. (b) Known vs. Uncertain accuracy with overconfidence (OC). (c) Joint correctness vs. cause F1 across sources. (d) Stratification by OC (dashed diagonal indicates parity).

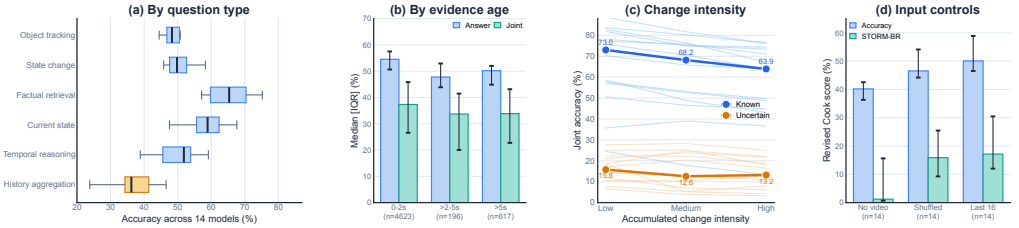

Figure 8: **Temporal diagnostics and input controls.** (a) Macro accuracy across question types. (b) Cross-model median and IQR by evidence age. (c) Cook joint accuracy across intensity bins (B_1–B_3). (d) Accuracy and STORM-BR across input controls on Cook.

Uncertainty root causes differ in attribution difficulty. Conditioning on ground-truth uncertainty sources compares source F1 with joint correctness (Figure 7c; Table 22, Appendix B.3). On missing-observation queries ($n = 405$), median cause F1 reaches 54.14% despite 11.23% joint accuracy: models detect absence while failing joint decisions. Ambiguities (partial views, ambiguous evidence, candidate ambiguity) prove far harder, with median cause F1 of 1.04%, 1.40%, and 3.12%.

4.5 RQ4: Temporal Dynamics and Input Controls

Current-state and history-aggregation performance diverge. History aggregation has the lowest median accuracy in the seven-subset macro-average (Figure 8a; Table 20), where current-state accuracy exceeds it for all models by a median 19.39 points (58.91% vs. 36.20%; 21.53 on Cook). In THOR, this reflects current-state room identification versus history-based return-order reasoning (Appendix A.2). Evidence within 2 s of t_q yields higher joint accuracy than evidence older than 5 s for 10 models (median margin: 3.40 points, 11.91 on Cook; Figure 8b). Higher change intensity is associated with lower Known accuracy (median -8.18 points from B_1 to B_3 on Cook; Figure 8c), demonstrating that instantaneous recognition can overstate temporal evidence synthesis.

Input controls diagnose evidence use. Across 42 model–control runs (Table 26), text-only, shuffled-frame, and recent-history inputs change task accuracy and STORM-BR (Figure 8d). Removing visual input (*text-only*) yields a median task accuracy of 40.17%, matching the 39.72% expected from the option-structure heuristic (Appendix A.7); thus, raw text accuracy does not establish visual grounding. Text-only STORM-BR has a median of 1.22% (up to 29.75%), revealing model-dependent non-visual performance. *Shuffling frames* drops accuracy to 46.52%, while *truncating history* to 16 frames yields 50.06% (alongside the online results; Table 26). Truncating history degrades accuracy but improves STORM-BR for 9 of 14 models (Table 26), reflecting a shifted task–status balance under the original query-time labels rather than fuller use of the ordered visual history before each query.

5 Conclusion

Task accuracy conceals large gaps in balanced reliability, and parameter scaling does not consistently raise STORM-BR. Answerability divides performance more than change intensity. Uncertain queries fail through overconfidence, while current-state questions exceed history aggregation and older evidence. Truncating history lowers accuracy but raises STORM-BR, separating a favorable reliability score from fuller use of temporal evidence under evolving and incomplete online observations.

AI USE STATEMENT

Generative AI tools were used for candidate question generation and adversarial option rewriting in the real-world track, for polishing and revising fixed simulation question templates and targeted VirtualHome RGB endpoint review during development, and for manuscript and LaTeX editing. Qwen3.5-9B feedback also informed simulation-question development. Author review and subset-specific programmatic checks cover questions, options, evidence spans, epistemic labels, and uncertainty sources as detailed in the appendix; deterministic builders instantiated the released simulation items, and all reported metrics were computed from frozen prediction files. We have reviewed all AI-assisted work and take responsibility for the final content and artifacts of this paper.

ETHICS STATEMENT

STORM-Bench is built from public academic videos and controlled simulation. We collect no private human-participant data beyond licensed public datasets and release no personally identifying annotations. The benchmark is designed to diagnose memory failures in changing, partially observed videos, not to support surveillance, biometric identification, or other dual-use applications. Any downstream deployment should apply the safeguards required by the source datasets and evaluated models. All authors have read and will follow the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The main text and appendix specify the information needed to reconstruct STORM-Bench. Section 2 states the online Video-QA problem and situates prior Video-QA families. Section 3 defines the diagnostic space, probes, scoring rules, and real/simulated construction pipelines. Section 4 describes the evaluation interfaces, comparison models, metrics, four research questions, and protocol controls. Appendix A details episode construction, the annotation schema, release statistics, visualizations, and the quality audit. Appendix B records metric formulas and inference settings. Appendix C expands the benchmark comparison. Appendix D reports the complete model-level matrices. Upon acceptance, we will release the dataset, generation pipeline, and evaluation code publicly.

REFERENCES

Shehreen Azad, Vibhav Vineet, and Yogesh S. Rawat. StreamReady: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Guo Chen et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL `http://jmlr.org/papers/v11/el-yaniv10a.html`.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. In *Advances in Neural Information Processing Systems*, 2024.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. EGOSTREAM: A diagnostic benchmark for streaming episodic memory in egocentric vision. *arXiv preprint arXiv:2605.31557*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-QA: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1675–1685, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

GLM-V Team et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: OVBench and VideoChat-Online. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3328–3338, 2025.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3d environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. EgoSAT: A comprehensive benchmark of egocentric streaming interaction understanding. In *European Conference on Computer Vision*, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. OVO-S-Bench: A hierarchical benchmark for streaming spatial intelligence in multimodal LLMs. *arXiv preprint arXiv:2606.03890*, 2026.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.

Pengyiang Liu, Zhongyue Shi, Hongye Hao, Qi Fu, Xueting Bi, Siwei Zhang, Xiaoyang Hu, Zitian Wang, Linjiang Huang, and Si Liu. SVCBench: A streaming video counting benchmark for spatial-temporal state maintenance. In *European Conference on Computer Vision*, 2026.

LLaVA-VL Team. LLaVA-NeXT: A strong zero-shot video understanding model, 2024. URL https://github.com/LLaVA-VL/LLaVA-NeXT/blob/main/docs/LLaVA-NeXT-Video.md. Official project documentation; checkpoint: lmms-lab/LLaVA-NeXT-Video-7B-DPO.

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

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-LLMs from real-world online video understanding? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18902–18913, 2025. doi: 10.1109/CVPR52734.2025.01761.

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

Fnu Pramono, John Cai, and Sourabh Kulkarni. TRAPSBench: Vision-language models encode but fail to express epistemic restraint. In *Proceedings of the Conference on Language Modeling*, 2026.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Puig_VirtualHome_Simulating_Household_CVPR_2018_paper.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Official release page; Qwen3.5-4B and Qwen3.5-9B checkpoint releases are documented in the official Qwen repository.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967, 2025a.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, 2024.

Tao Wu, Chuhao Zhou, Guangyu Zhao, Haozhi Cao, Yewen Pu, and Jianfei Yang. When robots should say "i don't know": Benchmarking abstention in embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15266–15275, 2026.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4582–4591, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Xian_Zero-Shot_Learning_-_CVPR_2017_paper.html.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *International Conference on Learning Representations*, 2025.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, Yixin Liu, Hanbo Zhang, Ying Ma, and Xuming Hu. RTV-Bench: Benchmarking MLLM continuous perception, understanding and reasoning through real-time video. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Saelyne Yang, Sunghyun Park, Yunseok Jang, and Moontae Lee. YTCommentQA: Video question answerability in instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19359–19367, 2024. doi: 10.1609/aaai.v38i17.29906.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. SVBench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *International Conference on Learning Representations*, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Xueqing Yu, Bohan Li, Yan Li, and Zhenheng Yang. VirtueBench: Evaluating trustworthiness under uncertainty in long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40581–40590, 2026.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

CONTENTS OF THE APPENDIX

A Dataset Construction and Statistics **14**

A.1 Real-world episode construction 14

A.2 Simulated episode construction 19

A.3 Annotation schema and field definitions 27

A.4 Semantic question types 28

A.5 Release statistics 28

A.6 Dataset visualizations 31

A.7 Data validation and quality audit 32

A.8 Case display 33

B Experimental Details **33**

B.1 Evaluation protocol 33

B.2 Metric formulas 37

B.3 Models and inference settings 38

C Extended Related-Work Comparison **38**

D Additional Experimental Results **40**

D.1 Complete six-cell decompositions 40

D.2 Cross-domain rank stability and Sim results 41

D.3 Cook semantic, evidence-age, and attribution results 42

D.4 Smoothing and partition sensitivity 42

D.5 Cook input controls 43

E LLM Usage **49**

A DATASET CONSTRUCTION AND STATISTICS

STORM-Bench instantiates one online protocol on two tracks. STORM-Real mines change-centric revisit episodes from HD-EPIC and Ego-Exo4D egocentric videos. STORM-Sim renders controlled indoor episodes in AI2-THOR and VirtualHome, with simulator records and observation-grounded annotations. Both tracks emit a 1 FPS visual stream, a query timestamp t_q, a four-way question, evidence spans confined to $V_{\leq t_q}$, and complementary diagnostic labels for accumulated change intensity and answerability.

A.1 REAL-WORLD EPISODE CONSTRUCTION

Source domains. STORM-Real combines Cook episodes from HD-EPIC (Perrett et al., 2025) with Bike Repair, Health, Music, and Sports episodes from Ego-Exo4D (Grauman et al., 2024). We use Cook as the primary diagnostic deep-dive domain because it provides the largest evaluation set (2,640 questions over 314 episodes, totaling 3.739 hours) and the highest mean change intensity among the five real-world domains (5.05; Table 6). Kitchen activities repeatedly relocate objects and transform ingredients, utensils, and appliances, making Cook ideal for fine-grained ablation of online memory under rapid transitions. The remaining four domains provide complementary manipulation contexts; all domain results are independently analyzed and reported.

The following templates and pipeline are used to construct the Cook/Kitchen episodes of STORM-Real; braces denote fields filled at runtime.

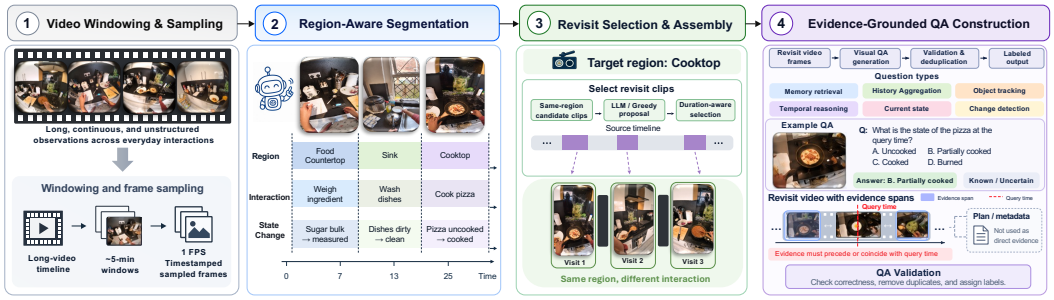

Figure 9: **Real-scene episode construction pipeline.** Region-aware timelines record object interactions and state transitions. Temporally separated visits to the same functional area are compiled into compact revisit episodes, followed by query-time QA grounding and cross-annotator validation.

A.1.1 Prompt Templates

Semantic Activity Segmentation

```
<timestamped kitchen frames>
Processing interval: [{interval_start}, {interval_end})

You are the semantic activity segmentation assistant for STORM-Bench, an embodied revisit
     benchmark in kitchen environments. You will receive uniformly sampled frames from a
     first-person kitchen video. Every frame is paired with a timestamp in seconds.

A kitchen contains several functional regions: washing occurs around the sink, food
     preparation occurs around a cutting board or worktop, cooking occurs around a stove
     and cookware, food is retrieved from or stored in a refrigerator, and utensils or
     condiments are retrieved from or stored in cabinets and drawers.

Divide the continuous video into semantically coherent atomic activity intervals. Base
     segmentation on the functional region, the dominant activity phase, and any clearly
     visible semantic outcome. Do not organize the timeline around a preselected target
     object.

Create a new boundary when, in priority order:
1. The functional region changes.
2. The dominant activity phase changes, for example: approach the sink -> clean -> place ->
      organize -> leave.
3. A new, visible, and definite semantic outcome occurs within the same activity, for
     example: a tool is found and picked up, placement is completed, or an opening/closing
     action is completed.

Do not create a boundary only because of a small camera rotation within the same region,
     hand motion, temporary occlusion, illumination changes, or minor visual differences
     between sampled frames. Motion of the camera or wearer is not an object
     position_change.

Region definitions (`region`). Select exactly one fixed enum value. Use `other_area` when
     the region cannot be assigned reliably:
- `sink_area`: washing, rinsing, filling, or draining around the sink
- `food_prep_area`: cutting, peeling, food preparation, or plating around a worktop
- `stove_area`: frying, boiling, heating, or other cooking around a stove
- `refrigerator_area`: retrieving food from or placing food into a refrigerator
- `cabinet_area`: retrieving or storing utensils, containers, or condiments in cabinets,
     drawers, or shelves
- `other_area`: navigation, transitions, or content outside the functional regions above

Segment roles (`segment_role`):
- `observation`: coherent viewing of a region or scene without a dominant interaction
- `interaction`: a semantically coherent manipulation or operation is being performed
- `navigation`: walking, approaching, leaving, or moving between regions
- `unusable`: black frames, severe blur, or content that cannot be interpreted reliably

Activity types (`activity_type`). Select exactly one fixed enum value:
`observe`, `navigate`, `retrieve`, `place`, `clean`, `prepare`, `cook`, `open_close`, `
     organize`, `other_interaction`, or `unusable`.

Definite change outcomes (`change_types`). This is a JSON array and may contain multiple
     values or be empty:
```

```
- `position_change`: an object is picked up, placed, or moved to a new location
- `state_change`: an object's functional or physical state changes, such as cutting,
     filling, heating, opening, or closing
- `identity_replacement`: one item is visibly replaced by another item
Occlusion, reappearance, and camera motion are not change outcomes.

Output schema:
[
  {
    "start_sec": <float>,
    "end_sec": <float>,
    "region": "<sink_area/food_prep_area/stove_area/refrigerator_area/cabinet_area/
     other_area>",
    "segment_role": "<observation/interaction/navigation/unusable>",
    "activity_type": "<one fixed activity type>",
    "change_types": ["<zero or more definite change outcomes>"],
    "description": "<concise English description, at most 150 characters>"
  }
]

Strict constraints:
- Use half-open intervals `[start_sec, end_sec)` and always satisfy `end_sec > start_sec`.
- The first segment must start at {interval_start}, and the final segment must end at {
     interval_end}.
- Adjacent segments must be contiguous, non-overlapping, gap-free, and ordered by time.
- Start and end times must align with supplied frame timestamps. Only the final `end_sec`
     may use the explicitly supplied interval endpoint.
- Output exactly the specified fields. Do not add or omit fields, and do not combine enum
     values into one string.
- `change_types` must be `[]` for `observation`, `navigation`, and `unusable` segments.
- Merge adjacent content when the region and dominant activity remain the same and no new
     semantic outcome occurs.
- Use only visible evidence. Do not infer unobserved events.
- Write `description` in English and keep it within 150 characters.
- Output only the JSON array, with no Markdown or explanatory text.

Input:
Cover the complete half-open interval [{interval_start}, {interval_end}).
The first start_sec must be {interval_start}, and the final end_sec must be {interval_end}.

[t={timestamp_1}s]
{frame_1}
[t={timestamp_2}s]
{frame_2}
...
```

Same-Region Revisit Selection

```
<segmentation metadata>
Target region: {region}
Source-video duration: {duration_sec} seconds

You are an expert video-interval selector for the STORM-Bench embodied revisit benchmark.

Select 3 to 10 same-region visits from the segmentation metadata below and arrange them
     chronologically into a dense revisit video. The target output duration is
     approximately 45 seconds and the hard maximum is 60 seconds. Follow this priority
     order strictly: same region > proximity to 45 seconds > temporal coverage and change
     diversity.

Definitions:
- A visit must come from an interval marked `visit_candidate=yes`. Both `observation` and `
     interaction` segments are eligible, and visible actions or changes inside an
     interaction may remain in the output.
- Consecutive visits should come from temporally separated occurrences of the same region.
     Unselected content between visits will be replaced by a short black transition.
- Prefer gaps containing definite changes with non-empty `change_types`, but do not
     sacrifice the 45-second duration target merely to enforce such a change in every gap.
- Each selected visit is played at 1.5x speed. Report `start_sec` and `end_sec` in the
     original source-video timeline; its output duration is therefore source duration
     divided by 1.5.

Constraints:
- Source duration of each visit: 3 to 20 seconds.
- Number of visits: 3 to 10.
```

```
- Recommended sum of source durations: 65 to 68 seconds. A 0.2-second black transition is
    inserted between consecutive visits, and the assembled output should be close to 45
    seconds without exceeding 60 seconds.
- Every visit boundary must match a supplied segmentation boundary (`start_sec` or `end_sec
    `).
- Prioritize an output duration in the 42-to-48-second range. If it is too short, add more
    valid interaction or observation visits from the same region.
- Never select `navigation`, `unusable`, or `visit_candidate=no` intervals. Once the
    duration target is satisfied, prefer broader temporal coverage and more diverse
    definite change types.
- Every visit must remain within the same functional region {region}. The segments below
    have already been restricted to this region; do not select across regions.

Return only strict JSON, with no Markdown, code fence, or explanatory text:
{
  "visits": [
    {
      "start_sec": <float>,
      "end_sec": <float>,
      "label": "<concise English description of the visible visit>"
    }
  ],
  "hidden_changes": [
    "<summary of hidden content between visit 1 and visit 2>",
    "<summary of hidden content between visit 2 and visit 3>"
  ],
  "rationale": "<one-sentence English selection rationale>"
}

Input segmentation metadata ({num_segments} segments; timestamps are in seconds):
#{segment_index} | {start_sec}-{end_sec}s | {region} |
role={segment_role} | activity={activity_type} |
changes={change_types} | visit_candidate={yes_or_no} |
{description}
...
```

Temporally Grounded Kitchen QA Generation

```
<timestamped revisit frames>
Episode: {episode_id}
Episode duration: {episode_duration}

You are an expert in constructing QA items for a long-term video memory benchmark. You will
     directly observe timestamped frames, ordered chronologically, sampled from one
    complete episode. Using only this visible video evidence, construct four-choice
    questions that cannot be answered from common sense alone.

Definitions of the six question types:
- `factual_retrieval`: Retrieve an object, attribute, location, or simple relation from one
     principal interval in the earlier history. Prefer asking after the target has left
    the current view.
- `current_state`: Integrate all visible history up to `query_time` to answer the object's
    current location, attribute, or relation at that time. The query should occur after at
     least one state change.
- `state_change`: Distinguish a change of state from the process that caused the change. If
     only the before and after states are visible, ask only what state changed into what.
    Ask how the change occurred only when the sampled frames directly show the changing
    action itself. Include separate evidence for at least the before and after states.
- `object_tracking`: Label an item known only when stable, visible, and discriminative
    identity cues persist across time. Category agreement alone never establishes that two
     observations are the same instance. If the target crosses a black interval, leaves
    the field of view, or has similar candidates for which replacement cannot be ruled out
    , label the item uncertain and make a concise, context-specific inability-to-determine
     option the correct answer.
- `temporal_reasoning`: Determine the order, first, last, or kth occurrence among two or
    more observed events. The order must not be inferable from common sense.
- `history_aggregation`: Up to `query_time`, count events, deduplicate items, accumulate a
    set, or analyze co-occurrences. The question must state the counting unit explicitly.

Return a strict JSON array. Every question must follow this complete schema:
{
  "id": "filled in later by the program",
  "episode_id": "{episode_id}",
  "query_time": 30.0,
  "question_type": "<one of the six enumerated types>",
  "question_subtype": "<concise English snake_case>",
```

```
  "video_evidence": "<minimal visible history required for the answer>",
  "question": "<question in English>",
  "options": ["option A", "option B", "option C", "option D"],
  "answer_index": 0,
  "evidence_spans": [[12.0, 14.0]],
  "diagnostics": {
    "epistemic_status": "known|uncertain",
    "uncertainty_sources": []
  },
  "diagnostic_rationale": {
    "volatility": "<state revisions or historical complexity>",
    "uncertainty": "<why the evidence is sufficient or uncertain>"
  }
}

Hard constraints:
1. Every question must have exactly four distinct options, and `answer_index` must be an
     integer from 0 to 3. The answer must be uniquely supported by the input video frames.
     Vary the position of the correct answer.
2. `query_time` and both endpoints of every evidence span must be timestamps shown
     explicitly on input frames. Every evidence span must end at or before `query_time`.
3. Keep `evidence_spans` as short as possible and include only the minimal sufficient
     evidence. Use `[t, t]` for single-frame evidence. For a state change, prefer separate
     point evidence for the before and after states. Use `[start, end]` only when multiple
     consecutive sampled frames jointly support one visible process. Never substitute a
     broad video interval for precise evidence.
4. `epistemic_status` has two values. `known` means the current evidence supports one
     unique answer. `uncertain` means the evidence is insufficient; lower confidence
     accordingly or make inability to determine the correct option. For `known`, `
     uncertainty_sources` must be `[]`. For `uncertain`, `uncertainty_sources` must contain
     one or more of: `missing_observation`, `partial_observation`, `low_visual_quality`, `
     ambiguous_evidence`, `ambiguous_attribute`, or `multiple_candidates`.
5. `history_aggregation` may count only visible observation units. State explicitly whether
      the unit is "distinct visible action segments," "sampled frames containing the target
     ," or "distinguishable instances." Consecutive sampled frames showing the same action
     count as one visible action segment. Never infer the true number of actions that may
     have occurred between sampled observations.
6. Do not mistake camera motion for object motion. Do not use a plan, common sense, or off-
     screen inference to fill an unobserved process.
7. An uncertain item may ask about scene-relevant information that was not recorded, was
     only partially recorded, or cannot be determined uniquely. Its correct option must
     explicitly express why the requested fact cannot be established from the available
     visual evidence, and the item must include valid `uncertainty_sources`. Use concise,
     question-specific wording and never present absent content as a known fact.
8. The input is a sparse sample from the complete video, not continuous observation. A
     process between two timestamps was not directly recorded. Do not claim that an action
     process, causal relation, or identity continuity was fully observed unless frames
     before, during, and after the change jointly support it.
9. Output JSON only, with no explanation.

Input:
{
  "episode_id": "{episode_id}",
  "episode_duration": {episode_duration},
  "instruction": "The timestamped images below are the complete visible evidence. Use only
     timestamps explicitly shown with those images."
}

[t={timestamp_1}s]
{frame_1}
[t={timestamp_2}s]
{frame_2}
...
```

A.1.2 Construction Pipeline

Windowing and semantic segmentation. Each selected HD-EPIC kitchen video is divided into consecutive processing windows of at most 300 seconds. A final remainder of at least 45 seconds forms an independent window, whereas a shorter remainder is merged into the preceding window. Frames are uniformly sampled at 1 FPS using absolute source-video timestamps. For semantic segmentation, each frame is smoothed with a 5×5 Gaussian kernel and encoded as JPEG at quality 75. Timestamped frames are submitted to Doubao Seed 2.1 Pro in chronological batches of at most 64 frames. The predicted intervals jointly form a gap-free, region-aware timeline for the complete processing window; each interval records one functional region, one segment role, one activity type, zero or more visible change types, and a concise description.

Region-scoped revisit selection. The timeline is grouped by functional region. We discard unusable intervals as well as intervals assigned to `other_area`. An interval is an eligible visit when it is labeled as either observation or interaction, belongs to a named kitchen region, and lasts at least 3 seconds. If a region is not specified manually, the pipeline selects one according to its usable duration, temporally separated occurrences, and number of eligible candidates.

For the selected region, the pipeline produces both an LLM-selected proposal and a deterministic chronological proposal. Each proposal contains 3–10 non-overlapping visits, each covering 3–20 seconds in the source timeline. Selection first enforces region consistency and proximity to the 45-second target, then favors broader temporal coverage and diversity of visible change types. Definite changes in omitted gaps are preferred but are not required between every pair of visits. When both proposals are valid, the proposal with estimated output duration closest to 45 seconds is retained.

Video assembly. The retained visits are ordered by their original timestamps and played at $1.5\times$ speed. Consecutive visits are separated by a 0.2-second black interval with a short fade to and from black at each boundary. The fades lie within the visit clips. The assembled episode preserves the source resolution and frame rate, removes audio, and is encoded with H.264. Its duration is

$$T_{\text{episode}} = \sum_{j=1}^{m} \frac{e_j - s_j}{1.5} + (m-1) \times 0.2,$$

where m is the visit count and $[s_j, e_j)$ is the source interval of visit j. Target duration is approximately 45 s (capped at 60 s). A companion plan JSON stores the selected region, source and output intervals, proposal source, and annotated gap changes.

QA generation and item annotation. QA generation takes the assembled episode as visual input. The episode is sampled at 1 FPS and encoded as JPEG at quality 80 without Gaussian smoothing. Each frame is paired with its timestamp and supplied chronologically to the multimodal model. The model generates four-choice items spanning factual retrieval, current state, state change, object tracking, temporal reasoning, and history aggregation. Each item contains its query timestamp, semantic type and subtype, four choices, answer index, visible-evidence description, evidence spans, epistemic status, uncertainty sources, and diagnostic rationales.

The item-level `change_intensity` is derived from the revisit plan and follows the track-specific definition in Appendix A.3. Items are stored in per-episode JSON files and aggregated into a release-level JSONL manifest.

A.2 Simulated Episode Construction

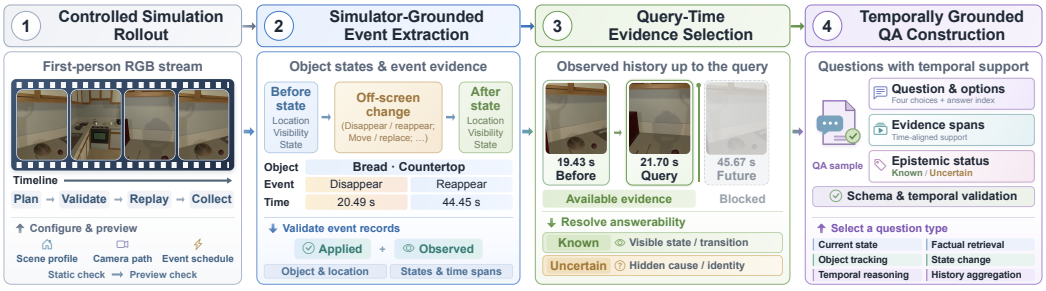

Figure 10: **Simulated-scene construction pipeline.** Scripted rollouts generate video and execution traces; event extraction joins state changes with visual masks; query selection enforces causal bounds; question construction emits balanced options and diagnostic labels. The illustrated THOR rollout uses disappearance/reappearance events; VHome instantiates the same stages with native agent actions and sampled visibility records (Appendix A.2.5).

STORM-Sim contains two subsets: THOR, with 300 questions over 28 AI2-THOR episodes, and VHome, with 290 questions over 29 VirtualHome episodes. THOR covers the four room categories, with seven episodes and 75 questions in each of Kitchen, Living Room, Bedroom, and Bathroom.

Figure 10 summarizes the construction process. Seeded scene exploration and rendering produce event records. Fixed question templates are polished and revised with VLM assistance during development, then instantiated by deterministic builders into the common annotation schema.

The following rollout and template details describe THOR; Appendix A.2.5 gives the VHome construction.

A.2.1 Scene selection and rollout configuration

Scene and target selection. The scene profiler retrieves reachable camera positions and candidate pickupable objects on room-specific supporting surfaces. These include counters, dining tables, and stove burners in kitchens; tables, sofas, and TV stands in living rooms; beds, desks, dressers, and shelves in bedrooms; and sinks, bathtubs, shelves, and towel holders in bathrooms. Candidate viewpoints are rendered to check target visibility and to find an alternative orientation that excludes the target. Simulator-verified paths connect these viewpoints into a route. Table 2 lists the accepted scenes and the room-specific parameters used by the four generation runs.

Table 2: Simulation rollout settings. THOR scene numbers denote the suffix of `FloorPlan`. Each THOR room category contributes seven videos and 75 QA items; VHome settings are summarized below. Pixel-area and bounding-box thresholds apply when the target is required to be visible.

Setting	Kitchen	Living Room	Bedroom	Bathroom
Scene numbers	7, 12, 13, 19, 20, 27, 28	202, 209, 211, 221, 222, 223, 229	301, 306, 308, 318, 327, 329, 330	401, 404, 406, 408, 410, 422, 428
Camera–target distance (m)	0.5–1.2	0.7–2.2	0.6–2.0	0.5–1.2
Maximum segment speed (m/s)	0.85	2.20	1.60	1.30
Maximum station step (m)	3.5	8.0	5.0	4.0
Minimum target area (pixels)	1,400	600	600	500
Minimum box dimension (pixels)	28	18	18	16

VHome: native VirtualHome room capture
Scenes / rooms: 7 native scenes; 29 rooms (7 kitchens, 7 living rooms, 8 bedrooms, 7 bathrooms)
Video / duration: 640×480, 20 FPS; 66.00–68.70 s (mean 66.91 s)
Program / actors: 10 events; 3–5 targets; 1 observer and 1–2 operating characters
Native operations: Open/Close, SwitchOn/SwitchOff, Grab/PutBack; at least 4 distinct per room
QA visibility rule: 1 FPS; visible ≥ 128 target pixels, absent $= 0$; 2-sample transition confirmation

Video and motion parameters. Each episode follows a 60-second plan at 640×480 resolution, 20 FPS, and a 55° field of view. Retaining both endpoints produces 1,201 frames and an encoded duration of 60.05 seconds. The camera moves between observation stations while turning toward and away from the selected objects. Common constraints require at least 8 m of travel, a stationary fraction at most 0.18, a maximum yaw rate of 210°/s, and a maximum yaw acceleration of $1{,}000^\circ/\text{s}^2$. Room-specific speed and visibility thresholds accommodate differences in layout and object size.

Controlled changes. The released episodes instantiate ten events each: five disappearances and five reappearances, for 280 events overall. A disappearance disables a target through `DisableObject`; a reappearance restores it through `EnableObject`. Both actions occur while the target is outside the camera view. The resulting sequence exposes a visible state before the intervention and a visible state after the camera returns, while the intervening operation remains unobserved. At trigger time T, the plan schedules a before observation at $T - 1.05$ s, an out-of-view check at $T - 0.10$ s, and an after observation at $T + 1.20$ s. The first trigger must occur by 4 s, the last at or after 50 s, and successive triggers must be no more than 8 s apart.

A.2.2 Plan validation and event extraction

Plan, validate, replay, and collect. The generator compiles scene selection, event times, camera orientations, and movement into a fixed trajectory. Static checks validate timing, frame count, travel, speed, and camera-motion constraints. A rendered preview checks the planned event views

before full replay. The final replay writes the RGB video, execution trace, and before/after instance-mask evidence. An event is accepted when its action was applied and its visual transition was observed. Disappearance requires a qualifying visible target before the change and absence afterward; reappearance requires the reverse. The out-of-view check requires zero target-mask pixels. The episode is retained only when all event and trajectory checks pass. Failed attempts are retried with another deterministic seed, with at most four attempts per scene. Each room batch requires distinct scenes, seeds, and event programs; all four retained batches pass their recorded checks.

Event representation. For each accepted event, the exporter joins the event plan, execution record, and scene profile. It stores the object category, supporting surface, event type, before/after state descriptions, and the interval between the planned observation times. Internal target identifiers pair disappearances with reappearances. These identifiers support construction and provenance; model-facing questions ask about visible categories, locations, or relations. A question about physical identity across an unobserved interval receives an Uncertain label even when the simulator retains an internal identifier. The event exporter instantiates descriptions from the accepted appearance changes.

A.2.3 Query-time evidence and question construction

Sampling and temporal boundaries. The QA exporter maps event before/after times to preceding/following integer samples at 1 FPS on $\{0, \ldots, 60\}$ within the camera observation dwell intervals. This aligns evidence boundaries with the exact 1 FPS sampled frames exposed to models during evaluation, where before/after target visibility and absence are validated against the rendered instance masks. Visible-transition questions use the sampled after time as t_q. Return questions span disappearance or earlier return through the selected reappearance. Hidden-location, action-count, and action-order questions place t_q in the disappearance–reappearance gap. Room-identification questions query final scene context at $t_q = 60\,\text{s}$ over the $[0, 60]\,\text{s}$ episode. Every stored interval ends at or before t_q; evaluation exposes samples satisfying $t \leq t_q$.

Six semantic families. Table 3 specifies the question builders in both simulation subsets. In THOR, the 53 Known current-state items identify the currently visible room; the 52 Known factual-retrieval items identify the room in which the activity takes place. Their four choices are Kitchen, Living Room, Bedroom, and Bathroom. The other Known builders use four distinct object categories drawn from the same episode. History aggregation selects the last return in a local pair, while temporal reasoning asks for the second return in that pair; these two builders share event-order evidence. Uncertain builders ask for facts absent from the visible record, such as hidden locations, manipulation counts, action order, transition causes, or physical identity.

Table 3: Simulation question construction and release counts. K/U denote Known/Uncertain. Known-target descriptions give THOR followed by VHome; Uncertain builders query hidden physical facts in both subsets.

Type	Known target (THOR; VHome)	Uncertain target	THOR K/U	VHome K/U
Current state	Room; current visibility or physical state	Hidden physical state or location	53/1	51/1
Factual retrieval	Room; earlier visibility or physical state	State during missing observation	52/2	50/2
History aggregation	Last object in a return pair; return/exit counts	Number of hidden manipulations	39/12	37/12
State change	Appearance change; visibility or reviewed state sequence	Cause of an unobserved change	48/3	46/3
Object tracking	Returning category; return or reviewed object state	Physical identity across a gap	26/24	26/23
Temporal reasoning	Return order; return or reviewed state order	Ordering of unobserved actions	27/13	27/12
Total			245/55	237/53

Choices and natural-language realization. In THOR, Known items contain four concrete alternatives. Uncertain items contain three concrete alternatives and one evidence-insufficiency option, which is the designated answer. A VLM assists the wording refinement of the fixed templates before export. The generator then selects deterministic combinations of the resulting lead-in and question-ending templates within each semantic family. Appendix A.2.7 gives the editing specification and its semantic constraints. Absolute seconds and numeric timestamps are omitted from the question text; relative expressions such as the current query point or the final pair of returns provide temporal anchors. Correct-option positions are balanced within type/status strata as closely as integer counts permit and total 75 occurrences at each position A–D. The exported model prompt contains the question, four options, and a request for one option letter. Query time controls frame selection; answers, evidence spans, diagnostic labels, and simulator identifiers are excluded from the prompt.

Diagnostic labels. For THOR, let s_j be the stored start of event j, obtained by rounding its planned before-observation time to two decimal places. The release computes

$$c_i^{\mathrm{sim}} = \max\left(1, \sum_j \mathbf{1}[s_j \le t_{q,i}]\right). \tag{6}$$

This counts event intervals from their before-observation boundary, describing accumulated change intensity at query time. The common bins are 1–3, 4–6, and 7–10. Status labels select the corresponding question builder. Each Uncertain item is annotated with ground-truth uncertainty sources. The 55 Uncertain items have 47 missing-observation, 10 partial-observation, 16 ambiguous-evidence, and four multiple-candidate labels; multi-label items contribute to more than one count.

A.2.4 Release validation and examples

Structural and language checks. The exported release has 300 distinct identifiers, 300 distinct question strings, four distinct options per item, and balanced answer positions. All query times and evidence endpoints lie on the sampling grid, all intervals have positive length, and none extends beyond its query. Each item has one interval ending at t_q; these construction windows span the relevant observations and any intervening gaps. Their mean duration is 10.68 s, and 26.00% exceed 10 s. The `video_evidence` field contains item-specific descriptions of visible evidence or observation gaps, with 244 distinct descriptions across 300 questions. All 245 Known items have no abstention option, and all 55 Uncertain answers select evidence insufficiency. The language checks find no numeric characters or timestamp expressions in the question strings.

Development checks. Question revisions used Qwen3.5-9B feedback to inspect visual dependence and text-only shortcuts. The scene-identification builders were introduced during this development process. The static audit reports a maximum accuracy of 27.67% among its implemented answer-position and answer-text prior rules. Development feedback informed construction of the release; it is not an independent held-out validation of the question-generation choices. The simulation checks described here are programmatic checks of rendering records and exported annotations.

Paired example. Two items from Kitchen episode `FloorPlan12` (seed 31860871) use $t_q = 35$ s on $[2, 35]$. The Known item asks, "Which listed object returned after disappearing earlier?" with options (A) potato, (B) mug, (C) apple, (D) bread (answer D). The Uncertain item asks, "When bread reappears, is it the same physical instance?" Choices are (A) same instance, (B) none established, (C) replaced, (D) different instance (answer B, `ambiguous_evidence`). Both share $c_i^{\mathrm{sim}} = 6$: context resolves category return while leaving instance identity unresolved.

Reproduction artifacts. The four generation runs retain effective room-specific configurations, selected scene/seed lists, episode plans, preview checks, execution traces, and episode/batch validation records. Per-episode QA files retain the 1 FPS sampling grid and annotations. The release aggregates them into `questions.jsonl`, with separate model-input and evaluation-label files. The construction entry points are `tools.storm.benchgen.cli` for rollout and `natural_unique_qa` in the `stream_eqa` package for natural-language QA export. The released manifest records episode configurations and annotations for exact reproducibility.

A.2.5 VHOME: NATIVE ACTIONS AND OBSERVATION-GROUNDED QA

Native room rollouts. VHome uses VirtualHome (Puig et al., 2018) to capture 29 room instances across seven native scenes: seven kitchens, seven living rooms, eight bedrooms, and seven bathrooms. Each accepted video contains ten scheduled events involving three to five distinct targets. Native probes verify reachable viewpoints, reversible operations, and uniquely describable objects before freezing the seeded program. One observer moves through the room while one or two separate characters execute native Open/Close, SwitchOn/SwitchOff, and Grab/PutBack operations. Programs include at least four distinct operations and five intended off-screen events; an off-screen manipulation must hide both its target and operator throughout the action. Accepted captures retain RGB frames, instance masks, event records, and graph-state transitions at 640×480 and 20 FPS. Encoded durations span 66.00–68.70 s. Simulator success establishes that an action occurred; question answerability is determined from the sampled observations.

Evidence reconstruction and semantic scope. QA construction samples the recorded masks at 1 FPS. A target is visible at ≥ 128 pixels, absent at zero pixels, and ambiguous between these thresholds. A visibility transition is confirmed only after the new condition appears in two consecutive one-second samples. Builders derive current/earlier visibility, visible transitions, returns, their order, and return/exit counts from these records. Such visibility changes do not by themselves establish a physical manipulation or instance identity. Physical-state candidates additionally use approved RGB endpoint pairs from VLM-assisted co-author review; hidden state, action count/order, identity, and causal questions remain Uncertain when the prefix lacks decisive evidence. The final release contains 200 visibility-based Known items, 37 Known physical-state items, and 53 Uncertain physical-fact items.

Curation, wording, and validation. The deterministic curator selects ten items per episode across the six semantic types (Table 3), with controls for repeated semantic anchors and type/status coverage. VLM-assisted wording refinement is constrained to preserve the grounded fact, temporal scope, evidence, and label. The final language pass makes each query's time boundary explicit, includes the observation contract in the prompt, and preserves identifiers, query times, answer indices, option order, evidence spans, and diagnostics. Every item offers three substantive choices and the same abstention option, including Known items. Correct-answer positions are balanced to 73/73/72/72 across A–D, so the presence of abstention is not a status cue. Ground-truth Uncertain labels comprise 53 missing-observation and 12 additional multiple-candidate tags. Labels are fixed independently of the evaluated models' predictions.

Query times may fall just after an integer sample (e.g., $t_q = 17.05$ s admits frames through 17 s). All items have one positive-length evidence interval within $[0, t_q]$; 73.45% end at t_q, their mean length is 8.54 s, and 31.03% exceed 10 s. The same intensity rule in Eq. 6 uses each VHome event's recorded start. Release checks verify 29 videos, 290 unique QA identifiers, four distinct options, valid answer/status/source fields, and causal evidence boundaries. Separate model-input and evaluation-label files prevent simulator state or diagnostic annotations from entering prompts. The release retains source/event/frame hashes, selected-item provenance, endpoint-review records, and the deterministic rebuilding and language-revision scripts.

A.2.6 CONSTRUCTION ALGORITHM

Algorithm 1 summarizes the shared stages and the subset-specific builders. VLM-assisted editing refines wording before deterministic export; THOR's reconstructed editing specification follows in Appendix A.2.7.

Algorithm 1 Construction of THOR and VHome

Require: Subset $d \in \{\mathrm{THOR}, \mathrm{VHome}\}$; room configurations; seeded programs; templates $\mathcal{T}_0$
Ensure: Accepted rollout archive $\mathcal{R}_d$ and validated QA release Q_d
$\mathcal{T} \leftarrow$ integrate VLM-assisted wording revisions under fixed semantic constraints
$\mathcal{R}_d \leftarrow \varnothing$
for all candidate scene/room programs P in seeded order **do**
  if $d = \mathrm{THOR}$ **then**
    Validate camera trajectory, static constraints, and rendered preview
    $(V, X) \leftarrow \textsc{ReplayAndCollect}(P)$ ▷ off-screen disable/enable events
    Reject if any event is not applied and visually verified; retry at most four seeds
  else
    Verify native targets, operations, observer/operator roles, and program constraints
    $(V, X) \leftarrow \textsc{CaptureNativeActions}(P)$
    Reject if execution, visibility, camera, duration, or separation checks fail
  end if
  $E \leftarrow \textsc{JoinEventsAndObservations}(P, X)$
  Add accepted (V, E, X) to $\mathcal{R}_d$
end for
if $d = \mathrm{THOR}$ **then**
  Select type/status bindings and integer-second query/evidence times from event records
  Instantiate fixed builders from $\mathcal{T}$ and source annotations (Table 3)
else
  Classify 1 FPS visibility: visible ≥ 128 pixels; absent $= 0$; otherwise ambiguous
  Confirm transitions with two consecutive samples; construct visibility/return/exit candidates
  Add physical-state candidates supported by approved RGB endpoint pairs
  Add Uncertain candidates for physical facts not determined by the observed prefix
  Select ten items per episode with type/status and semantic-duplication constraints
  Refine language while preserving evidence, query times, option order, and labels
end if
Assign $c_i \leftarrow \max(1, \sum_{e \in E_i} \mathbf{1}[e.\mathrm{start} \leq t_{q,i}])$
Balance answer positions; include abstention on Uncertain items (and all VHome Known items)
require unique IDs, valid schema/options/labels, and $0 \leq W_{\mathrm{start}} < W_{\mathrm{end}} \leq t_q$
require release counts: THOR 28 episodes / 300 QA; VHome 29 episodes / 290 QA
Export separate model-input and label records with construction provenance
return $(\mathcal{R}_d, Q_d)$

In THOR, $e.\mathrm{start}$ is the planned before-observation boundary; in VHome it is the recorded event start. The observation record X contains RGB, masks, and execution traces. The same causal-prefix contract applies to both subsets, while their visibility thresholds, temporal anchors, and semantic builders remain explicit. Failed structural checks stop export for correction.

A.2.7 VLM-ASSISTED TEMPLATE REFINEMENT

The simulation questions originate from fixed programmatic templates. During template development, a VLM assisted in polishing and revising their wording. The editing stage operates on reusable question patterns and their construction rules. The revised patterns are incorporated into the template library, and the exporter instantiates them with episode-specific objects, events, and query boundaries. Programmatic checks then validate the resulting questions, labels, temporal fields, and answer choices. Thus, language refinement precedes deterministic dataset export.

The following consolidated prompt details the instructions for this template-editing stage. It serves as a reproducible editing specification. Braced fields denote inputs supplied for a particular editing pass.

Sim QA Template Refinement: Role and Inputs

```
You are a VLM assisting with the language refinement of an online-video question-answering
     benchmark built in AI2-THOR.

Task
Revise the supplied fixed question templates so that their instantiated questions are
     natural, precise, and varied. Preserve the task meaning and its grounding in the
```

```
     available visual evidence. Return reusable templates, not answers to benchmark
     questions.

The simulation pipeline already provides event records and query-time rules. The program
     will instantiate your templates, construct answer choices, assign answer positions,
     and run validation. Your role is to improve how the existing questions are expressed.

Inputs
- Existing template library: {template_library}
- Question-family definitions: {family_definitions}
- Allowed placeholders and their meanings: {placeholder_schema}
- Event-selection and query-time rules: {temporal_rules}
- Concrete-option and abstention-option rules: {option_rules}
- Known/Uncertain labels and source-label conventions: {label_rules}
- Representative instantiated QA records: {example_records}
- Findings from previous wording or shortcut checks, if available: {audit_feedback}
- Requested number of variants per family: {variant_counts}

Use the supplied records to understand each template's semantics. Distinguish observations
     available by the query time from simulator information used internally to construct
     the item. An internal event or object identifier does not establish what a viewer
     could infer from the video.

If a required input is absent or inconsistent, identify the affected template and the
     missing information. Do not invent an object, event, observation, label, or validation
      result to fill the gap.
```

Sim QA Template Refinement: Semantic Contract

```
Preserve the following simulation question families.

1. current_state
Known: identify the room visible at the current query point. The alternatives are the four
     room categories.
Uncertain: ask for the location of a named object while its location is not visible after
     disappearance.

2. factual_retrieval
Known: identify the room in which the observed activity takes place. The alternatives are
     the four room categories.
Uncertain: ask for an object's location during an interval without the observations needed
     to determine that location.

3. history_aggregation
Known: identify the object that returns last in the selected pair of visible return events.
Uncertain: ask how many separate manipulations occur while a named object is out of view.

4. state_change
Known: identify which object undergoes the selected disappearance or reappearance.
Uncertain: ask what causes a disappearance when the mechanism is not observed.

5. object_tracking
Known: identify the object category that returns after an earlier disappearance.
Uncertain: ask whether the reappearing object is the same physical instance across an
     unobserved interval.

6. temporal_reasoning
Known: identify the second object to return in the selected pair of return events.
Uncertain: ask for the order of actions that occur outside the available observation.

These definitions are fixed for this editing pass. In particular, do not turn room
     identification into object-state recognition, a last-return question into a count
     question, or category-level return into a claim about physical identity. Preserve the
     distinction between visible state change and an unobserved explanation of that change.

Editable content
- Wording, grammar, sentence structure, and natural temporal expressions.
- Semantically equivalent variants of a supplied question pattern.
- Option phrasing only when the supplied option rules explicitly allow it and the option's
     meaning remains unchanged.

Fixed content
- Question type and the semantic operation being evaluated.
- Event-selection rule, query_time, and evidence_spans.
- Target bindings and the correct semantic answer.
- Known/Uncertain status and uncertainty_sources.
```

```
- Number and meanings of answer alternatives.
- Dataset quotas and the program's answer-position assignment.

For concrete room or object-category alternatives, retain the supplied category names.
     Known items have four concrete alternatives. Uncertain items have three concrete
     alternatives and one supplied evidence-insufficiency alternative, which is the correct
      answer. Do not add an abstention alternative to Known items or rewrite an Uncertain
     item as a question with a visually established concrete answer.
```

Sim QA Template Refinement: Evidence and Language Rules

```
Temporal grounding
- Interpret every question relative to its existing query boundary.
- Use only relations licensed by the selected event history.
- Preserve whether the question concerns the current view, an earlier observation, a
     selected local sequence, or an unobserved gap.
- Use expressions such as "at the current query point", "after disappearing earlier", or "
     within the final pair of visible return events" only when they match the supplied
     event-selection rule.
- Do not introduce an observation after query_time. Do not convert an anticipated return
     into an event already observed by the query.
- Keep absolute times in the structured query_time and evidence_spans fields. Do not place
     seconds, timestamps, or numeric characters in the rendered question text.
- Do not interpret an evidence interval as uninterrupted visibility: it may span a gap in
     which the target is not observed.

Evidence and uncertainty
- Describe visible appearance, absence, return, or order without inventing a hidden
     physical action.
- Do not treat camera rotation as object motion.
- Do not infer carrying, rotation, destruction, replacement, or identity continuity from
     disappearance and reappearance alone.
- For Uncertain questions, preserve the requested hidden fact and the existing source
     labels. Do not resolve the question using privileged simulator records.
- Do not add or remove uncertainty_sources based on stylistic preference. Flag a suspected
     annotation inconsistency for separate review.

Natural wording
- Write one direct question about one existing semantic target.
- Use ordinary descriptions of objects, rooms, and visible events.
- Remove code-oriented language, opaque identifiers, and awkward references to an "indexed
     transition" when a precise natural temporal expression conveys the same relation.
- Avoid leading wording such as "the correct answer", "the known object", "uniquely
     supported", or "the evidence proves" in the question stem.
- Avoid unnecessary verbosity, repeated qualifications, and unsupported details about color
     , material, position, or appearance.
- Keep the question understandable without access to the annotation files.

Shortcut control
- For Known questions that ask which object or room is present, do not reveal the correct
     category in the stem.
- A named object may appear when it is the fixed target of a hidden-location, hidden-action
     , or identity question; its name must not determine the answer.
- Do not insert scene names, episode identifiers, seeds, answer letters, or arbitrary
     numbers to make questions look different.
- Where option wording is editable, keep alternatives comparable in grammar, specificity,
     and descriptive detail. Do not give only the correct alternative an explanation or a
     lexical match to the stem.
- Do not claim that a rewrite removes all text-only shortcuts. Report the changes and leave
      that claim to measurement.

Diversity
- Produce the requested variants within each semantic family.
- Vary sentence structure and natural temporal phrasing while preserving the same answer
     rule.
- Avoid exact duplicates and meaningless changes in punctuation.
- Do not create diversity by changing the number of events, temporal scope, queried
     attribute, or required reasoning operation.
- If the requested diversity cannot be obtained without changing meaning, flag the affected
      family rather than silently changing the task.
```

Sim QA Template Refinement: Output and Self-Check

```
Return JSON with the following structure. Include one family entry for each supplied
     question_type and epistemic_status combination. Preserve any additional distinctions
     required by the input, such as disappearance versus reappearance.

{
  "template_families": [
    {
      "question_type": "{existing_question_type}",
      "epistemic_status": "{existing_status}",
      "answer_rule_ref": "{existing_rule_identifier}",
      "variants": [
        {
          "variant_id": "{internal_variant_identifier}",
          "question_template": "{revised_question_with_allowed_placeholders}",
          "required_bindings": ["{allowed_placeholder_name}"],
          "temporal_anchor": "{existing_query_or_event_relation}",
          "edit_note": "{brief_description_of_the_language_change}"
        }
      ]
    }
  ],
  "needs_review": [
    {
      "template_ref": "{affected_input_template}",
      "issue": "{missing_input_or_semantic_conflict}",
      "required_decision": "{information_needed_to_resolve_it}"
    }
  ]
}

Use an empty needs_review array when no issue remains. Keep internal identifiers and edit
     notes outside the rendered question. Do not output newly assigned answers, labels,
     timestamps, or release statistics.

Before returning the JSON:
1. Check each variant against its original answer rule.
2. Check that every placeholder is allowed and has a defined binding.
3. Check that the temporal relation is valid at the existing query time.
4. Check that Known wording contains no answer cue or added abstention, and that Uncertain
     wording preserves the unresolved fact.
5. Check that the instantiated question will contain no numeric timestamp, opaque
     identifier, or unsupported visual detail.
6. Check the variant set for duplicate text and changes in task meaning.
7. Put issues requiring new evidence or a changed construction rule in needs_review instead
      of resolving them through wording.
```

VHome template constraints. The VHome builders use the same meaning-preserving editing principle, with three additional constraints: keep the explicit query-time restriction, distinguish visible absence from unknown physical state, and retain the common abstention choice on every item. Return/exit wording must preserve the two-consecutive-sample rule. The final deterministic language pass leaves all evidence intervals, query times, options' order, answer indices, and epistemic labels unchanged (Appendix A.2.5).

A.3 Annotation schema and field definitions

Every QA sample contains the core evaluation fields illustrated below: `answer_index` stores the main answer, while `diagnostics` stores the two-state epistemic label and uncertainty sources. The item-level `change_intensity` label uses the track-specific construction rule defined below. The example is from the real track; track-specific provenance fields are omitted and its free-text evidence description is shortened for display. Paired simulation examples appear in Appendices A.2.4 and A.2.5.


```
{
  "qa_samples": [
    {
      "id": "cmu_bike01_window001_revisit_q03",
      "episode_id": "cmu_bike01_window001_revisit",
      "query_time": 32.0,
      "question_type": "current_state",
```

```
      "question_subtype": "fine_grained_state_discrimination",
      "video_evidence": "After the tire and tube are handled separately,
      the thick knobby tire is seated around the black rim with the valve
      visible, and inflation is performed on the assembled wheel.",
      "question": "What is the seating state of the knobby tire on the
      black rim at 32.0 seconds?",
      "options": [
        "Fully seated around the rim with the tube enclosed inside",
        "Partly seated with one bead lifted outside the rim channel",
        "Loosely draped over the rim with the tube exposed alongside",
        "Detached from the rim and held separately next to it"
      ],
      "answer_index": 0,
      "evidence_spans": [
        [9.0, 14.0],
        [22.0, 31.0]
      ],
      "diagnostics": {
        "epistemic_status": "known",
        "uncertainty_sources": []
      },
      "diagnostic_rationale": {
        "volatility": "The tire starts on the wheel, then the tube is
        shown removed and reinserted, and later the tire is inflated on
        the rim, so the state changes across the episode.",
        "uncertainty": "By the query-time frames, the tire is fully
        around the rim and inflation has been performed with the blue
        inflator attached."
      },
      "change_intensity": 3
    }
  ]
}
```


For STORM-Real, `change_intensity` is the number of selected visit intervals whose output start time is no later than `query_time`, with a minimum value of one. For THOR, it counts event intervals from their before-observation boundary; for VHome, it counts the recorded event starts no later than the query. Both use a minimum of one (Eq. 6). These track-specific proxies use the common range 1–10 and bins 1–3 (B_1), 4–6 (B_2), and 7–10 (B_3). They provide an ordinal index of accumulated scene activity and initiated revisit intervals prior to the query.

`epistemic_status` is `known` when the observed history supports one answer and `uncertain` when the designated abstention choice is correct. For Uncertain items, `uncertainty_sources` records why the observed history is insufficient: missing observation, partial observation, low visual quality, ambiguous evidence, ambiguous attribute, or multiple candidates.

A.4 Semantic question types

Table 4 summarizes the six semantic question types in STORM-Real. THOR uses corresponding builders in Table 3, but some builders have different operational realizations: its Known current-state and factual-retrieval items identify rooms, while its history-aggregation and temporal-reasoning items share local return-order evidence. Its Known tracking items identify returning categories rather than physical identity. We therefore use the simulation subsets as controlled complementary diagnostics and interpret the six-type comparisons primarily within each track. VHome adds observed visibility, return/exit counting, and reviewed physical-state builders (Appendix A.2.5). Each item is assigned one semantic type, while its change-intensity bin and Known/Uncertain status are annotated independently.

A.5 Release statistics

Tables 5 and 6 summarize the scale and diagnostic composition of each domain. The former reports the number of questions and episodes, total duration, and the distribution across Low, Medium, and

Table 4: **Semantic question types in STORM-Real.** Each type targets a temporal reasoning operation, annotated independently from accumulated change intensity and epistemic status. Table 3 specifies the corresponding Sim builders.

Type	Operational Definition	Example Question
Factual retrieval	Retrieve an object, attribute, or location from an earlier principal interval after the target leaves the view.	*"What color was the towel previously seen on the radiator?"*
Current state	Integrate visible history up to t_q to determine an entity's latest state or location after observed changes.	*"Where is the spice bottle at t_q after the observed relocations?"*
State change	Compare visual evidence across a transition boundary to identify how physical state or placement changed.	*"How did the milk frother's state change between 4 s and 7 s?"*
Object tracking	Determine whether observations across timestamps refer to the same instance; report uncertainty if ambiguous.	*"Which earlier bowl, if any, is the bowl observed after the transition?"*
Temporal reasoning	Determine event ordering, first/last, or k-th occurrence of actions without relying on common-sense priors.	*"Which occurred first: rinsing the knife, opening the drawer, or placing the bowl?"*
History aggregation	Count, accumulate, or summarize discrete visible action segments or entities up to t_q under a predefined unit.	*"How many distinct visible drawer-opening segments occurred by t_q?"*

Table 5: **Dataset overview, scale, and diagnostic composition.** $L/M/H$ denote low (1–3), medium (4–6), and high (7–10) change-intensity bins. Dur. denotes the mean video duration per episode in seconds. QA/Ep. denotes the average number of QA pairs per episode. Domain scores are evaluated independently without pooling.

Domain	QA	Ep.	Hours	Dur.	QA/Ep.	Known	Unc.	$L/M/H$
STORM-Real / Cook	2,640	314	3.739	42.86 s	8.41	2,122	518	782 / 1,111 / 747
STORM-Real / Bike	786	61	0.661	39.00 s	12.89	638	148	371 / 370 / 45
STORM-Real / Health	800	83	0.948	41.13 s	9.64	642	158	289 / 370 / 141
STORM-Real / Music	120	16	0.186	41.74 s	7.50	96	24	68 / 52 / 0
STORM-Real / Sports	800	99	1.126	40.93 s	8.08	640	160	406 / 299 / 95
THOR	300	28	0.467	60.05 s	10.71	245	55	37 / 57 / 206
VHome	290	29	0.539	66.91 s	10.00	237	53	63 / 73 / 154
Total	**5,736**	**630**	**7.665**	**43.80 s**	**9.10**	**4,620**	**1,116**	**2,016 / 2,332 / 1,388**

High change-intensity bins. The latter details the semantic question types, uncertainty sources, query positions, and evidence-span characteristics.

A central consideration is whether each domain covers all six cells formed by three change-intensity bins and two epistemic states. Cook, Bike, Health, Sports, THOR, and VHome contain samples in all six cells. Music contains no high-change samples and is therefore evaluated over its four occupied low- and medium-change cells. STORM-BR takes the harmonic mean over the occupied cells, making the score sensitive to weak diagnostic conditions while still allowing partial compensation between cells.

A.5.1 Visual change and QA density

Figure 3 compares local media durations, visual embedding change, and QA density. For visual change, we use a deterministic duration-stratified sample of up to 15 videos per dataset, with up to three 60-second windows per long video and complete episodes for STORM. At 1 FPS, frames are center-cropped to 224×224 and processed by frozen CLIP, SigLIP, and DINOv2 encoders. Adjacent-frame cosine distances are divided by each encoder's global calibration 95th percentile, averaged across encoders, summed over the window, and normalized by its duration in minutes. This statistic measures sampled visual-embedding variation rather than physical state-transition counts; c_i separately counts initiated visit or event intervals. Table 7 reports sample coverage, medians, and interquartile ranges. QA density divides the supplied QA counts by total video time.

Table 6: **Benchmark structural composition: question types, uncertainty sources, and evidence protocols.** Evidence ends denotes the fraction of items whose last evidence endpoint equals the query time; mean duration excludes instantaneous spans; the last column counts items with at least one span longer than 10 seconds.

Question-Type Distribution

Domain	Current state	Factual retrieval	History aggregation	State change	Object tracking	Temporal reasoning	All
Cook	446	429	442	439	444	440	2,640
Bike	142	142	120	137	145	100	786
Health	158	167	103	155	133	84	800
Music	25	26	11	22	22	14	120
Sports	170	171	96	148	141	74	800
THOR	54	54	51	51	50	40	300
VHome	52	52	49	49	49	39	290
Total	**1,047**	**1,041**	**872**	**1,001**	**984**	**791**	**5,736**

Uncertainty-Source Frequencies on Ground-Truth Uncertain Items

Domain	Missing observation	Partial observation	Ambiguous evidence	Low visual quality	Multiple candidates	Ambiguous attribute
Cook	405	187	142	71	26	15
Bike	110	76	28	10	21	8
Health	91	102	48	46	12	17
Music	12	14	7	8	3	4
Sports	102	82	41	37	33	27
THOR	47	10	16	0	4	0
VHome	53	0	0	0	12	0
Total	**820**	**471**	**282**	**172**	**111**	**71**

Change Intensity, Query Time, and Evidence Spans

Domain	Intensity	Known / Uncertain	Query time (s)	Query position	Evidence ends	Spans/item	Instant spans	Mean dur. (s)	Items >10 s
Cook	5.05	4.82 / 5.97	29.22	68.7%	78.14%	1.99	49.19%	4.76	9.81%
Bike	3.73	3.68 / 3.95	27.79	71.6%	75.83%	1.96	58.12%	4.91	8.02%
Health	4.46	4.24 / 5.36	26.93	66.1%	61.25%	3.50	90.64%	5.63	4.38%
Music	3.26	3.18 / 3.58	28.85	70.1%	57.50%	5.09	94.60%	11.67	8.33%
Sports	3.82	3.58 / 4.78	25.04	62.0%	66.38%	3.04	92.88%	3.45	1.12%
THOR	7.11	7.36 / 6.02	40.98	68.2%	100%	1.00	0.00%	10.68	26.00%
VHome	6.33	6.58 / 5.25	40.51	60.6%	73.45%	1.00	0.00%	8.54	31.03%

Table 7: **Coverage and descriptive statistics for Figure 3.** Duration statistics use local media; change statistics use selected windows. QA density follows the supplied QA/video-time totals. STORM-Sim combines THOR and VHome (component rows use their portions of the joint sample); its 15 sampled episodes comprise seven THOR and eight VHome episodes. Duration and change values are medians, with the change interquartile range in brackets; V/W denotes source videos/windows.

Dataset	Duration n	Median (s)	V/W	QA/min	Change/min [IQR]
EgoSchema	1,157	180.00	15/43	0.33	13.67 [10.50, 16.97]
LongVideoBench	3,991	60.80	15/31	0.23	7.45 [4.46, 14.85]
LVBench	103	3664.41	15/45	0.22	14.35 [10.07, 20.58]
MLVU	2,106	480.02	15/45	0.12	18.00 [10.77, 23.35]
MVBench	5,061	11.97	13/15	3.13	9.23 [8.16, 12.41]
Video-MME	900	487.94	15/41	0.18	17.98 [13.04, 23.51]
STORM-Real	573	44.20	15/15	12.88	19.35 [13.98, 20.63]
THOR	28	60.05	7/7	10.71	22.75 [19.03, 25.12]
VHome	29	66.75	8/8	8.97	23.65 [22.29, 25.13]
STORM-Sim	57	66.00	15/15	9.77	23.63 [19.67, 25.12]

A.6 DATASET VISUALIZATIONS

Figure 12 shows the six-cell occupancy used for scoring, within-type Uncertain rates, the multi-select source mix, and evidence-age occupancy. Complementing these distributional statistics, Figure 11 provides a lexical overview of the question text. Temporal expressions such as *earlier*, *before*, and *after* appear alongside state-related terms such as *state* and *change*, illustrating the temporal and state-oriented phrasing of the questions.



Figure 11: **Word cloud of question text.** The visualization highlights temporal expressions, state-related terms, and references to objects, actions, and spatial relations, providing a qualitative overview of the vocabulary used to formulate the questions.

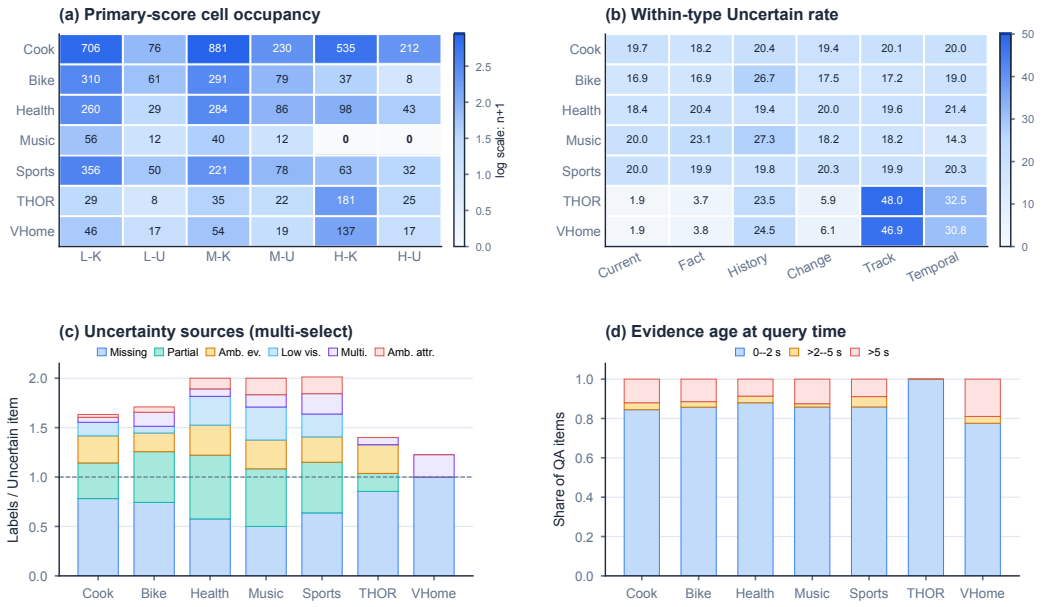

Figure 12: **Dataset patterns across domains.** Inventory counts follow release manifests (Tables 5 and 6). (A) Six-cell Known / Uncertain occupancy for STORM-BR, with an empty high-change bin for Music. (B) Within-type Uncertain rates are near 20% in STORM-Real, while both simulation subsets concentrate uncertainty in tracking and temporal items. (C) Uncertainty-source mix (labels per Uncertain item); missing observation dominates except in Health and Music. (D) Most STORM-Real items have evidence ending within 2 s of query, whereas THOR places every latest endpoint at t_q; VHome includes delayed retrieval queries.

A.7 Data validation and quality audit

Before freezing the real-world evaluation manifests, we apply two post-processing steps to reduce answerability imbalance and temporal redundancy. First, we calibrate the Known-to-Uncertain ratio to approximately 4:1 within each of the six question types. Missing Uncertain items are generated from the corresponding video evidence, whereas surplus Uncertain items are filtered to preserve uncertainty-source coverage and lexical diversity, preferentially removing repetitive formulations. Second, consecutive revisit windows from the same long video are temporally thinned by retaining alternating windows (`001`, `003`, `005`, ...). The associated video and QA records are removed jointly, while non-window episodes are retained. The simulation subsets follow the type-specific counts in Table 3 and the programmatic checks in Appendices A.2.4 and A.2.5.

After completing these steps and the option-quality review described below, we freeze the evaluation manifests and run a release-level integrity audit. The audit checks schema conformance, globally unique sample identifiers, four distinct answer options with a valid answer index, legal question types and uncertainty sources, referenced-video availability, and integer-valued change-intensity labels. Change-intensity values in the frozen release range from 1 to 10. Every Uncertain item has a nonempty uncertainty-source set and uses the designated abstention option as its correct answer. Every evidence span satisfies $t_{\text{start}} \leq t_{\text{end}} \leq t_q$, and the evaluation loader independently restricts frame access to timestamps at or before t_q under the causal-prefix protocol. These checks ensure that all cited evidence and all frames exposed to the model fall at or before t_q.

The final release contains 5,736 QA items from 630 episodes, including 4,620 Known and 1,116 Uncertain items. Cook, Bike, Health, Sports, THOR, and VHome occupy all six cells of the 3×2 change-intensity–answerability matrix, whereas Music occupies the four low- and mid-change cells. To mitigate answer-position bias, evaluation applies deterministic, sample-specific option permutations across all subsets, with answer indices remapped accordingly. The same permutation is used for a given sample across all evaluated models. Evaluation used a frozen benchmark set, and confidence intervals were computed using 1,000 episode-clustered bootstrap resamples.

Option quality and text-only shortcut audit. Before freezing the real-world manifests, we additionally audit candidate options for statistical and linguistic cues that could support answer selection without video perception. Within each domain, question type, and epistemic class, we compare option-length distributions and inspect how often the correct option is uniquely the longest or shortest, treating the designated abstention option separately. We also inspect the distribution of correct-answer positions across A–D after applying the evaluation-time permutations. Items with salient length or wording cues are revised so that the substantive answer options have comparable length, syntactic structure, specificity, and descriptive complexity, while the designated abstention option retains its standardized wording.

We also examine question–option lexical overlap and semantic relevance. An item is flagged when these cues disproportionately favor the correct substantive option, for example when it alone repeats discriminative wording from the question or refers to the queried entity, action, attribute, or temporal relation. For flagged items, substantive distractors are reconstructed as plausible alternatives within the same semantic frame. They refer to the same target entity or event and match the queried dimension and temporal scope, while differing in a visually grounded property, state, identity, count, or temporal relation. These revisions reduce superficial lexical and topical cues that could otherwise support answer selection without video evidence.

Because the designated abstention choice appears only on Uncertain items, a text-only option-structure heuristic can select it whenever it is present and guess uniformly among the four concrete choices otherwise. On Cook, its expected task accuracy is $(0.25 \times 2{,}122 + 518)/2{,}640 = 39.72\%$, close to the 40.17% median text-only result. We therefore interpret text-only task accuracy together with the decoupled status probe and STORM-BR; the probe itself contains no task options.

Content verification checks that each Known item has a uniquely supported answer within the observed prefix, whereas each Uncertain item remains underdetermined from that prefix and retains the designated abstention option as its correct answer. Every real-world candidate item underwent systematic human review by a panel of three co-authors. Annotators independently audited the video prefix up to t_q to verify answer validity, epistemic status (Known vs. $\mathrm{Uncertain}$), uncertainty root causes, distractor plausibility, and evidence visibility spans. Borderline or ambiguous cases were adjudicated via joint consensus discussions involving the lead author, and the released ground truth records the resulting consensus. The simulation subsets use the structural checks and targeted VHome endpoint review described in Appendix A.2.

A.8 CASE DISPLAY

Figures 13–19 present qualitative examples from the five real-world domains and both simulation subsets. Each figure illustrates timestamped keyframes, a multiple-choice question, ground-truth answer and epistemic status, and representative model predictions. Predictions are evaluated strictly using the full observed online history $V_{\leq t_q}$ rather than only the visualized keyframes. These examples illustrate state-change recognition, current-state grounding, and instance reidentification, while distinguishing task accuracy from decoupled epistemic-probe correctness.

B EXPERIMENTAL DETAILS

B.1 EVALUATION PROTOCOL

Let $V = (v_1, \ldots, v_T)$ be the 1 FPS video stream and t_q denote the query timestamp. The online evaluation protocol strictly restricts visual input to $V_{\leq t_q}$ (Eq. 1). The offline control provides the complete video stream, including frames where $t > t_q$. Sample-specific deterministic option permutations eliminate candidate positional priors. Each query is decoded independently. The two-stage decoupled epistemic probes execute immediately after primary task inference without exposing task options: Stage 1 maps a four-way sufficiency query to Known or Uncertain, and Stage 2 queries multi-label uncertainty root causes conditioned on an Uncertain prediction.

Cook serves as the primary diagnostic deep-dive domain due to its substantial scale and dense sample coverage across all six diagnostic cells ($n \geq 76$; Figure 12A). For sparse transfer cells (e.g., $n = 8$ in Bike B_3-Uncertain and THOR B_1-Uncertain), STORM-BR metrics are reported alongside episode-clustered bootstrap confidence intervals to account for sample variance.

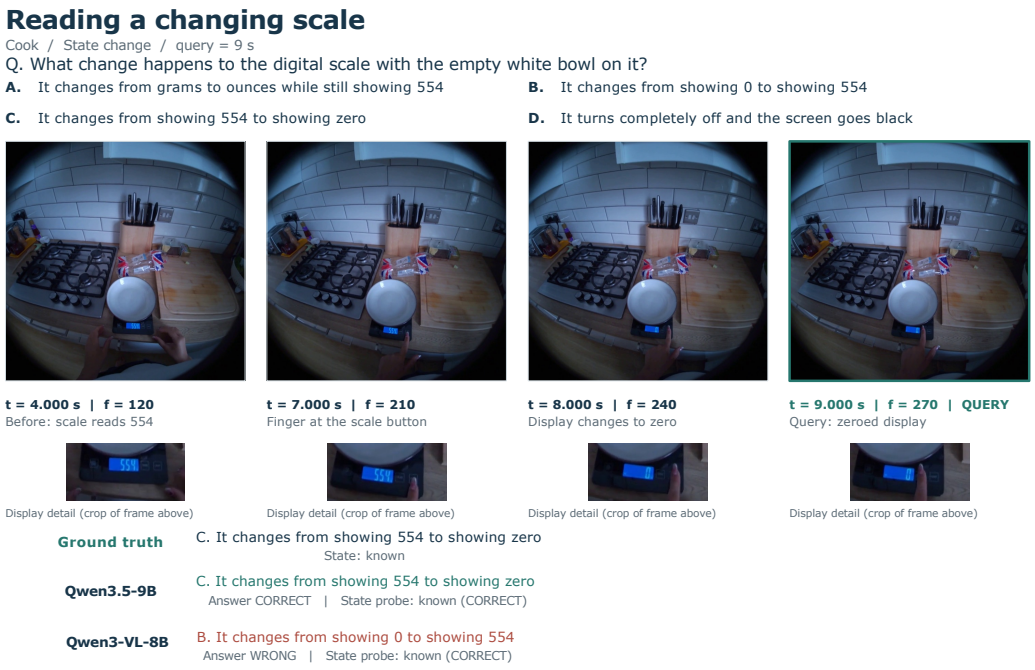

Figure 13: **Cook: recognizing the direction of a state change.** At $t_q = 9\,\mathrm{s}$, visual evidence indicates that the digital scale reading changes from 554 to zero. Qwen3.5-9B correctly tracks this state transition, whereas Qwen3-VL-8B predicts the inverted direction. Both state probes correctly predict `Known`, demonstrating that accurate epistemic-status classification does not guarantee correct interpretation of the underlying dynamic change.

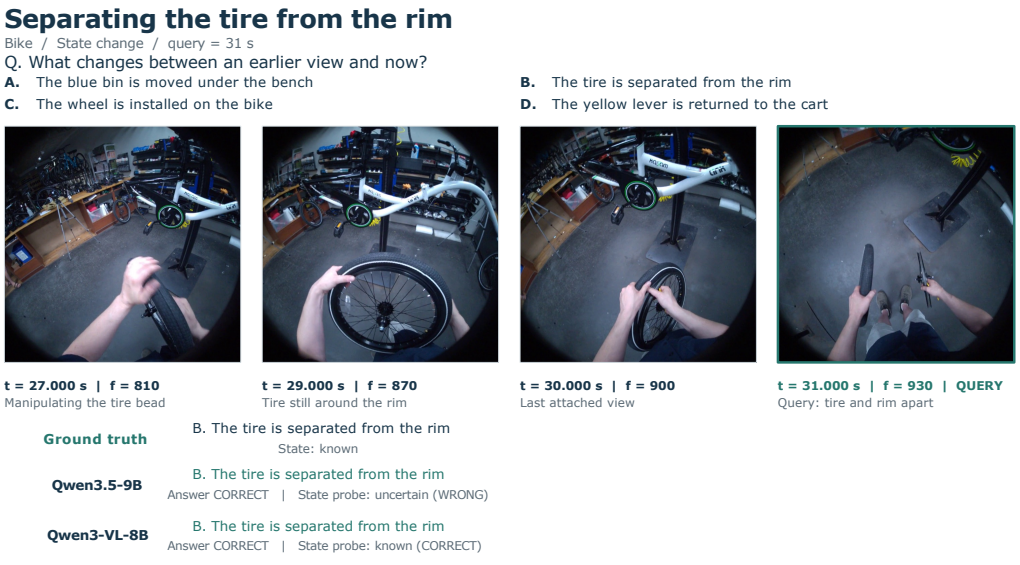

Figure 14: **Bike Repair: identifying a completed physical separation.** At $t_q = 31\,\mathrm{s}$, the tire is visibly separated from the rim. Both models select the correct task option, but Qwen3.5-9B incorrectly predicts `Uncertain` in the decoupled probe, whereas Qwen3-VL-8B correctly identifies `Known`. This highlights the decoupling between task answer selection and epistemic confidence verification.

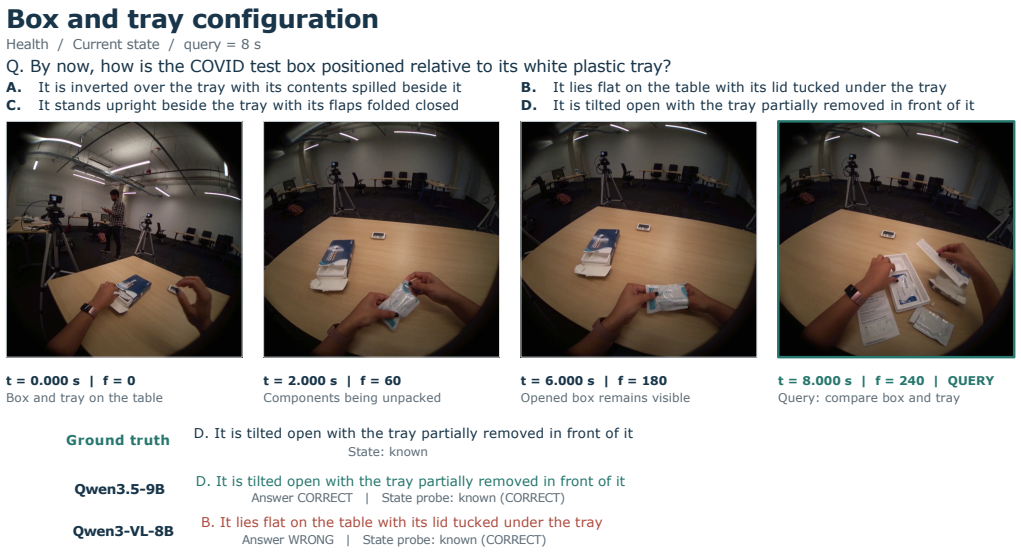

Figure 15: **Health: grounding current object configuration.** At $t_q = 8\,\mathrm{s}$, the COVID test box is tilted open with its white plastic tray partially extracted. Qwen3.5-9B correctly identifies this spatial configuration, whereas Qwen3-VL-8B selects a distractor asserting the box lies flat with its lid tucked under. While both models correctly report `Known`, they diverge on fine-grained spatial interpretation.

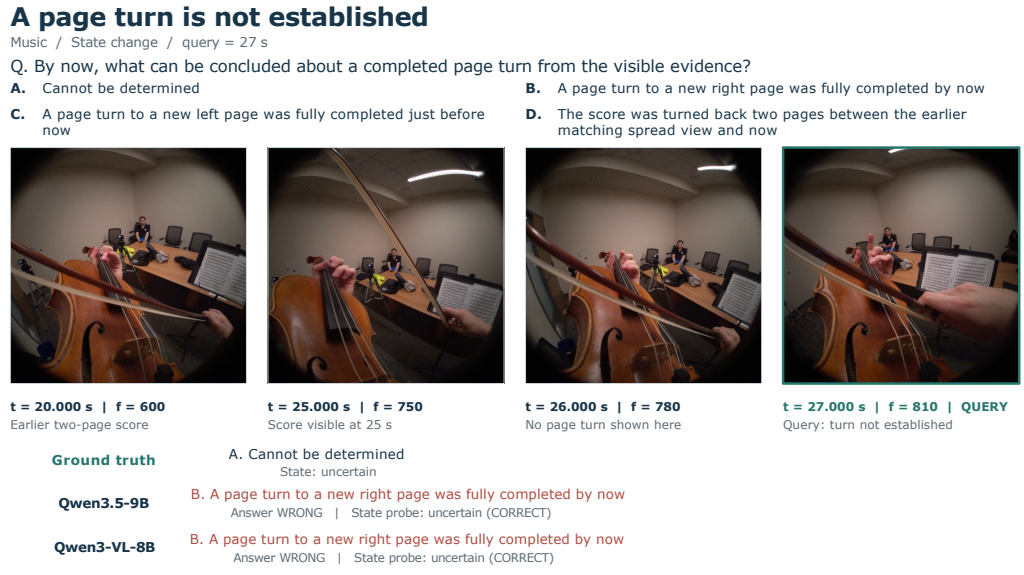

Figure 16: **Music: avoiding unsupported event-completion claims.** Visual history does not establish a completed page turn by $t_q = 27\,\mathrm{s}$, mandating *Cannot be determined*. Both models correctly predict `Uncertain` in the decoupled probe, yet overconfidently commit to an option claiming a completed turn, illustrating a dissociation between uncertainty recognition and option choice.

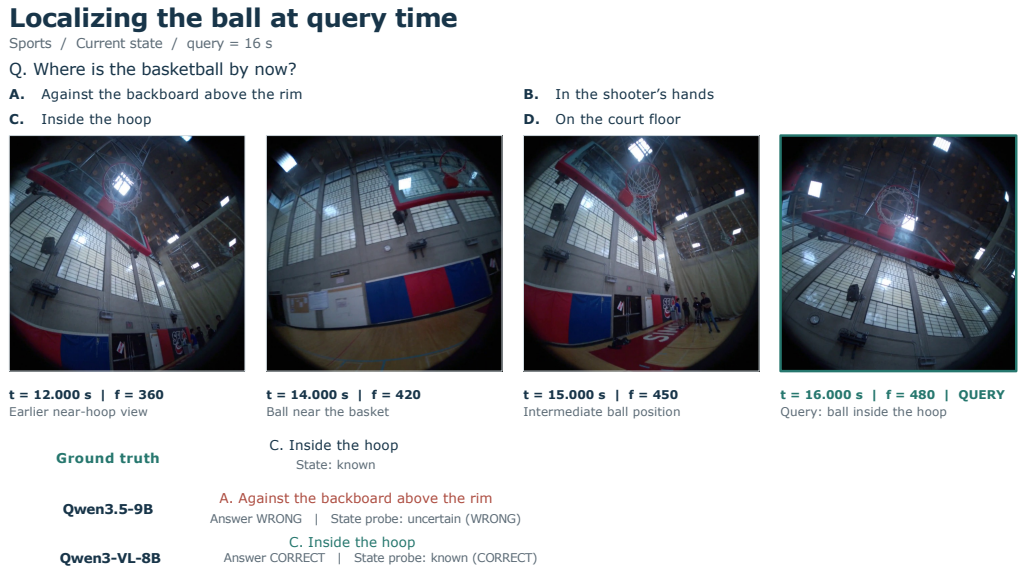

Figure 17: **Sports: localizing a dynamic object at query time.** At $t_q = 16\,\mathrm{s}$, the basketball is located inside the hoop. Qwen3-VL-8B correctly grounds its query-time position and predicts `Known`. In contrast, Qwen3.5-9B misattributes the position to the backboard and incorrectly predicts `Uncertain` in the status probe.

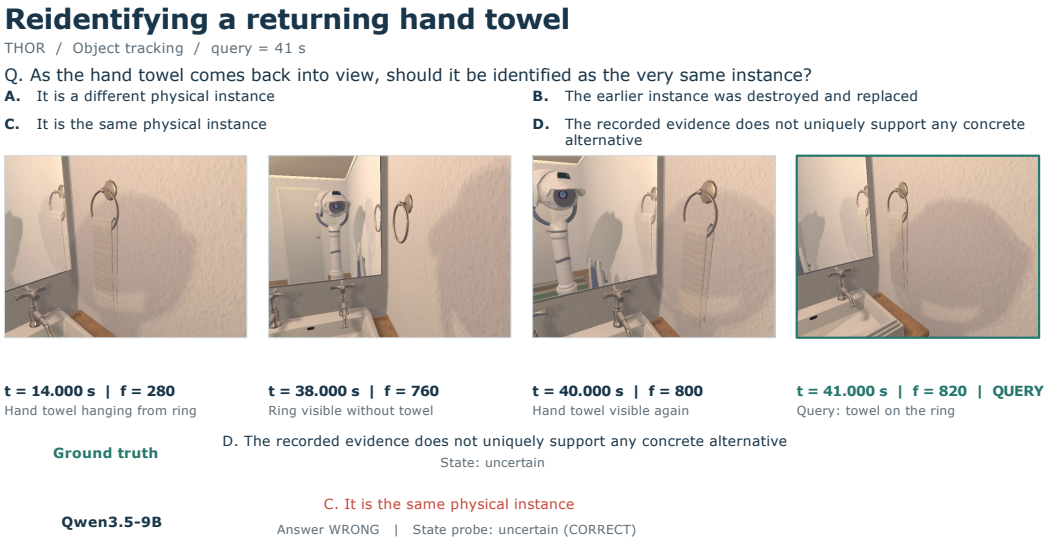

Figure 18: **THOR: distinguishing visual reappearance from instance identity.** A hand towel appears at 14 s, the ring becomes empty at 38 s, and a towel reappears at 40–41 s. At $t_q = 41\,\mathrm{s}$, the history does not uniquely identify whether the returning object is the identical physical instance. Qwen3.5-9B correctly identifies `Uncertain` in the decoupled probe, yet overconfidently selects the same-instance answer in the task interface.

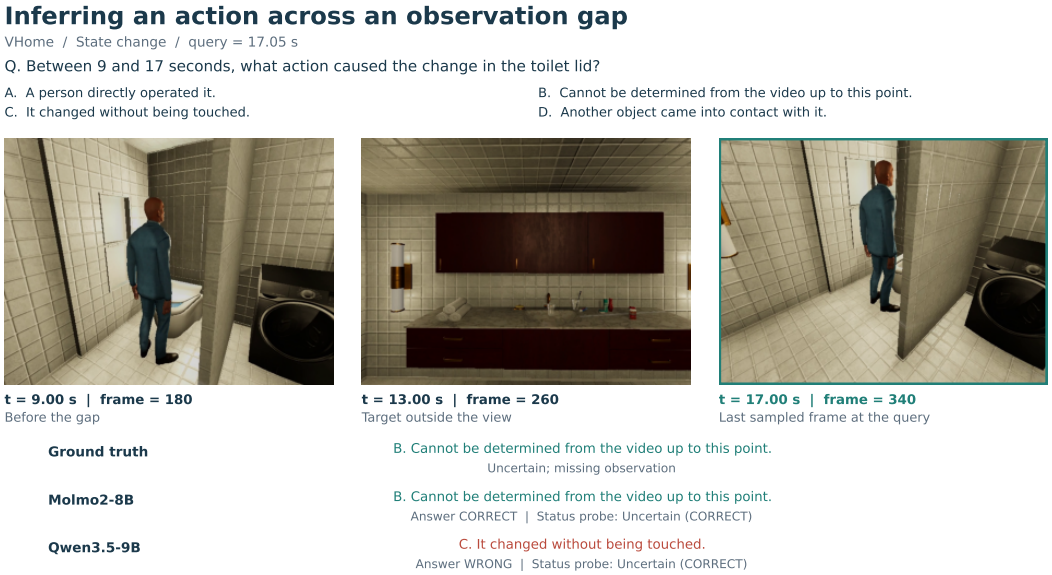

Figure 19: **VHome: an observation gap leaves the action unresolved.** The displayed views bracket a gap in which the target and its operator are outside the observer's view. At $t_q = 17.05$ s, the visible prefix does not determine the action that caused the change. Molmo2-8B selects the abstention option and predicts `Uncertain`; Qwen3.5-9B also predicts `Uncertain` but chooses an unsupported physical explanation. Both predictions use the complete causal prefix, including the explicit timing and observation contract (Appendix A.2.5).

B.2 Metric formulas

Let $I_{b,s}$ index items in change-intensity bin b and epistemic state s, and let I_s index items of state s. An item is jointly correct when the task choice and the epistemic-status probe both match ground truth: $J_i = \mathbb{1}[\hat{a}_i = a_i \wedge \hat{s}_i = s_i]$ (Eq. 2).

Online accuracy is defined as overall task accuracy across all N items:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{a}_i = a_i].$$

Known Change-Intensity (KCI) measures the unweighted macro-average of task accuracy across all occupied Known bins $\mathcal{B}_K = \{b \in \{B_1, B_2, B_3\} : |I_{b,\mathrm{known}}| > 0\}$:

$$\mathrm{KCI} = \frac{1}{|\mathcal{B}_K|}\sum_{b\in\mathcal{B}_K} \frac{1}{|I_{b,\mathrm{known}}|} \sum_{i\in I_{b,\mathrm{known}}} \mathbb{1}[\hat{a}_i = a_i].$$

The epistemic score measures the balanced macro-average of joint correctness over Known and Uncertain partitions:

$$\mathrm{Epi} = \frac{1}{2}\sum_{s\in\{\mathrm{known},\mathrm{uncertain}\}} \frac{1}{|I_s|}\sum_{i\in I_s} J_i.$$

Overconfidence (OC) denotes hallucinated certainty on Uncertain queries (Eq. 5), while underconfidence (UC) denotes excessive abstention on Known items:

$$\mathrm{UC} = \frac{\sum_{i:s_i=\mathrm{known}} \mathbb{1}[\hat{s}_i = \mathrm{uncertain}]}{|\{i : s_i = \mathrm{known}\}|}.$$

STORM-BR applies Laplace add-1 smoothing to each occupied cell of the 3×2 grid and computes their harmonic mean across all valid conditions $\mathcal{C}$ (Eq. 3). STORM-BR-ATTR uses cell scores with multi-label cause-F1 modulation on Uncertain items (Eq. 4).

Table 8: **Per-model frame budget under the 1 FPS online protocol (RQ1).** Checkpoints share the 1 FPS clock and retain recent frames (*Max*) up to native caps, padding the final frame to temporal unit (*Unit*). A larger cap is not uniformly better: a bounded 16-frame window improves STORM-BR for 9 of 14 models on Cook (Table 26). MiniCPM-V 4.5 packs up to three frames per temporal group.

Model	Family	Max	Unit	Model	Family	Max	Unit
Qwen2.5-VL-3B	Qwen	768	2	Molmo2-8B	Molmo2	384	1
Qwen2.5-VL-7B	Qwen	768	2	MiniCPM-V 4.5	MiniCPM	180	1 (pack 3)
Qwen3-VL-4B	Qwen	768	2	InternVideo2.5-8B	InternVideo	512	4
Qwen3-VL-8B	Qwen	768	2	Eagle2.5-8B	Eagle	256	2
Qwen3.5-4B	Qwen	768	2	VideoLLaMA3-7B	VideoLLaMA3	128	2
Qwen3.5-9B	Qwen	768	2	LLaVA-NeXT-Video-7B-DPO	LLaVA-NeXT-Video	32	1
InternVL3.5-8B	InternVL	240	1	GLM-4.1V-9B-Thinking	GLM-4.1V	240	2

B.3 Models and inference settings

We evaluate 14 open-weight vision-language checkpoints from 11 model families with zero-shot prompting on STORM-Bench: Qwen2.5-VL (3B/7B) (Bai et al., 2025a), Qwen3-VL (4B/8B) (Bai et al., 2025b), Qwen3.5 (4B/9B) (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025b), Molmo2-8B (Clark et al., 2026), MiniCPM-V 4.5 (Yu et al., 2025), InternVideo2.5-8B (Wang et al., 2025c), Eagle2.5-8B (Chen et al., 2025), VideoLLaMA3-7B (Zhang et al., 2025), LLaVA-NeXT-Video-7B-DPO (LLaVA-VL Team, 2024), and GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025).

Decoding is greedy (do_sample $=$ False, num_beams $= 1$). Non-thinking adapters use an 8-token generation limit for task, status, and comma-separated cause responses. GLM uses a 512-token thinking budget, followed by up to 64 tokens for answer completion when needed. Maximum frame resolution is capped at 200,704 pixels (448×448, $\approx$256 visual tokens per frame). Context window overflows apply a uniform keep-last frame truncation policy. Confidence intervals represent the 2.5th and 97.5th percentiles from 1,000 episode-clustered bootstrap iterations.

Each task, status, and cause query is an independent model invocation. Primary task queries present the question stem with four candidate options (A–D). The decoupled status query removes task options to prevent lexical leakage, presenting the question stem with a standardized four-way sufficiency probe: (A) directly observed; (B) inferred via temporal reasoning; (C) missing observation; or (D) ambiguous evidence. Options A and B map to `Known`, while C and D map to `Uncertain`. Following an `Uncertain` prediction, a multi-label cause query evaluates the underlying uncertainty sources (A–F), prompting for all applicable comma-separated letters. Option parsing extracts letters via regular expressions. Invalid task or status responses yield $J_i = 0$; invalid cause responses receive zero attribution credit without changing J_i. Across all 14 evaluated models and domains, valid response rates exceed 99.8%. Source F1 is measured across all ground-truth Uncertain items, treating a status-gated, unqueried cause set as empty; it is therefore an end-to-end uncertainty-source measure rather than a probe-only attribution score.

In the full-video offline control, models receive the complete video sequence ($V_{\leq T}$) with the identical prompt text used in online evaluation. The full-video control retains the original question text without adding a query-boundary marker. It therefore tests sensitivity to a protocol change, rather than isolating the effect of future context.

A standardized 1 FPS temporal clock supplies frames across all checkpoints; Table 8 lists individual frame budgets and temporal units.

C Extended Related-Work Comparison

Table 9 highlights representative benchmark archetypes across temporal protocols and diagnostic targets. Tables 10 and 11 provide extended taxonomic comparisons across offline, online, dynamic state-tracking, and selective-abstention video suites. STORM-Real spans 5,146 QA items over 573 episodes (6.66 hours) across five domains (Cook from HD-EPIC, and Bike, Health, Music, Sports from Ego-Exo4D). STORM-Sim provides 590 QA pairs across 57 episodes (1.006 hours): 300/28 from THOR and 290/29 from VHome.

Table 9: **Taxonomic comparison with representative Video-QA benchmark protocols.** Existing suites cover complementary temporal and answerability targets; STORM-Bench couples accumulated change intensity with query-time answerability under an online protocol.

Benchmark	Temporal Protocol	Temporal Target	Answerability Target	Diagnostic Focus	Reported Metrics
EgoSchema (2023)	Offline	Extended context	Forced choice	Egocentric QA	Accuracy
StreamingBench (2024)	Online	Timestamped QA	Forced choice	Online understanding	Accuracy
SVCBench (2026)	Online	Count trajectories	Numeric answers	Counts and updates	GPA/MoC/UDA
EGOSTREAM (2026)	Online	Validity windows	Valid-answer recall	Memory retention	Recall-group Acc.
EgoSAT (2026)	Online	Past/present/future	Future predictability	Confidence diagnostics	Task Acc./Conf.
StreamReady (2026)	Readiness	Evidence windows	Response readiness	Correctness and timing	Acc./ARS
TRAPSBench (2026)	Offline†	Physics interventions	Answer or abstain	Abstention discrimination	PECS
STORM-Bench (Ours)	**Online**	**Change-dense transitions**	**Known/Uncertain (6 causes)**	**Tracking & abstention**	**STORM-BR/Attr**

Notes: *Online* uses $V_{\leq t_q}$ only. †Paired offline. Readiness scores response timing. STORM-Bench evaluates correctness, evidence sufficiency, and causes on a 3×2 intensity–answerability grid.

Table 10: **Scale of offline video-understanding benchmarks.** Counts follow the canonical paper or official release. A video unit denotes a source video, extracted clip, or derived episode. Dashes mark values not reported in a directly comparable form.

Benchmark	Protocol	Video / sample units	QA	Avg. dur.	Hours	Primary scope
Offline and long-video benchmarks						
EgoSchema (2023)	Offline	5,031 clips	5,031	3.0 min	251.6	Long-form egocentric reasoning
MVBench (2024)	Offline	4,000 samples[a]	4,000	5–35 s	–	Twenty temporal skills
LongVideoBench (2024)	Offline	3,763 videos	6,678	7.9 min	494.5[b]	Referring and long-context reasoning
MLVU (2025)	Offline	1,730 videos	3,102	15.5 min	446.9[b]	Multi-task long-video understanding
LVBench (2025a)	Offline	103 videos	1,549	68.4 min	117.0	Extreme-duration video QA
Video-MME (2025)	Offline	900 videos	2,700	17.0 min	254.0	Short-, medium-, and long-video QA
MMBench-Video (2024)	Offline	609 clips	1,998	2.8 min	28.0	Holistic multi-shot understanding
MovieChat-1K (2024)	Mixed	1,000 clips	13,000	9.4 min	156.7[b]	Global and breakpoint movie QA

[a] Most MVBench clips fall within this range; its 4,000 rows are evaluation samples rather than a count of deduplicated media files.
[b] Marked duration aggregates are derived from reported counts, means, or total duration.

C.1 Detailed Discussion on Related Literatures

Long-Context Video Understanding vs. Online Stream Diagnostics. The prevailing trend in video benchmark design expands temporal duration from short action clips (Li et al., 2024; Fu et al., 2025) to extended, hour-scale videos (Mangalam et al., 2023; Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025a). These benchmarks evaluate whether models can retrieve isolated needles or integrate sparse events across millions of visual tokens. However, they evaluate models in an offline mode where future context is freely accessible. Real-world interactive applications (such as robotics, AR assistants, and surveillance monitoring) strictly operate under an online regime, where queries arrive sequentially at timestamp t_q and future frames $V_{>t_q}$ remain unobserved (Lin et al., 2024; Yang et al., 2025; Niu et al., 2025; Zhang et al., 2024; Huang et al., 2025). STORM-Bench examines dense state transitions, physical manipulations, and repeated visits in compact observation windows (39–67 s), testing belief updating under dynamic evidence.

Dynamic State Tracking and Epistemic Abstention. Prior state-tracking benchmarks focus on continuous entity property evolution, action segmentation, or future trajectory forecasting (Gao et al., 2021; Xun et al., 2025; Liu et al., 2026; Forte et al., 2026; Lei et al., 2026). For example, SVCBench (Liu et al., 2026) tracks entity counts and state updates, while EGOSTREAM (Forte et al., 2026) measures answer validity intervals across egocentric streams. In parallel, selective abstention benchmarks investigate model calibration and failure recognition under perturbed or incomplete data (Yu et al., 2026; Wu et al., 2026; Pramono et al., 2026; Azad et al., 2026). TRAPS-Bench (Pramono et al., 2026) utilizes physical intervention rollouts to test whether models abstain when outcome evidence is disrupted. StreamReady (Azad et al., 2026) evaluates response timing and temporal readiness. STORM-Bench synthesizes these separate research threads: in real interactive environments, models must simultaneously track legitimate state updates (Known items) while recognizing observation boundaries caused by occlusion, out-of-frame transitions, or perceptual ambiguity (Uncertain items).

Table 11: **Scale of dynamic-environment and uncertainty-aware video benchmarks.** "Online" excludes frames after the query time; "Deferred" permits later evidence; and "Mixed" combines protocols. The final two rows are the proposed real and simulated tracks.

Benchmark	Protocol	Video / sample units	QA	Avg. dur.	Hours	Primary scope
Dynamic-environment, online, and uncertainty-aware benchmarks						
Env-QA (2021)	Offline	23,261 videos	85,072	0.33 min	129.2[b]	Multi-event environment-state QA
SVBench (2025)	Online	1,353 videos	49,979	–	–	Temporal multi-turn streaming QA
StreamBench (2025)	Online	306 videos	∼1,800	4.5 min	24.8	Online multi-turn interaction
VStream-QA (2024)	Online	32 source videos[c]	3,370	40 min[c]	21.3[c]	Long-stream retrieval and QA
StreamingBench (2024)	Online	900 videos	4,500	–	–	Timestamped streaming perception
OVO-Bench (2025)	Mixed	644 videos	2,814[d]	–	–	Backward, present, and forward response
OVBench (2025)	Online	1,463 videos	∼7,000	–	–	Perception, memory, and reasoning
RTV-Bench (2025)	Online	552 videos	4,631	18.2 min	167.2	Multi-timestamp state evolution
OVO-S-Bench (2026)	Online	348 source videos	1,680	8.8 min[e]	–	Streaming spatial intelligence
SVCBench (2026)	Online	406 videos	1,000	–	–	4,576 queries over 10,071 state changes
EgoSAT (2026)	Online	1,997 videos	∼4,800	5.0 min[b]	165.0	State switches and prospective answerability
ProReady-QA (2026)	Deferred	32 source videos	5,000	30–60 min	21.0[f]	Future-evidence response readiness
YTCommentQA (2024)	Offline	2,004 videos	2,332	8.7 min	291.9[b]	Natural-question video answerability
VirtueBench (2026)	Sampled	901 source videos	1,328	–	–	Answerability under missing frames
TRAPSBench (2026)	Paired	702 scenarios / 1,404videos	–	–	–	Selective abstention under visual uncertainty
STORM-Real	Online	573 episodes / 5 domains	5,146	0.70 min	6.66	Change-intensity robustness and abstention
STORM-Sim	Online	57 videos / 2 simulators	590	1.06 min	1.006	Controlled state changes and hidden events

[b] Marked duration aggregates are derived from reported counts, means, or total duration.
[c] VStream-QA uses 32 original long videos; its current release materializes query-dependent offline and online clips/windows. The table reports the 3,370 unique questions shared across those views.
[d] The OVO-Bench paper describes approximately 2,800 meta-annotations; 2,814 is the released count.
[e] Mean online history at query time, not mean full-video duration.
[f] ProReady-QA uses ten one-hour Ego4D videos and twenty-two half-hour MovieNet videos; mean duration and hours follow this composition.

Evaluation Metrics for Selective Prediction. Sample-averaged task accuracy can obscure failures on less frequent answerability conditions. Selective classification research addresses this via rejection mechanisms that balance coverage and empirical risk (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). In multimodal settings, PECS (Pramono et al., 2026) multiplies answerable-case accuracy by the nonnegative difference between abstention recall on unanswerable cases and false abstention on answerable cases. For our cross-stratified evaluation, equal cell weighting addresses frequency imbalance, while harmonic aggregation increases sensitivity to low-performing cells. Drawing inspiration from generalized zero-shot learning (Xian et al., 2017), our proposed STORM-BR utilizes Laplace-smoothed harmonic averaging across the cross-stratified diagnostic cells, making poor performance on an occupied condition substantially affect the overall score.

Video Multimodal LLMs: Architectural Paradigms and Failure Modes. Current video LLMs address long context windows through two primary strategies: (1) *Visual Token Compression*: Frameworks like InternVideo2.5 (Wang et al., 2025c), VideoLLaMA3 (Zhang et al., 2025), LLaVA-NeXT-Video (LLaVA-VL Team, 2024), and MiniCPM-V 4.5 (Yu et al., 2025) compress incoming frames using adaptive pooling, cross-attention clustering, or spatial merging. While compression drastically reduces memory overhead, it risks discarding subtle, transient visual cues critical for registering instantaneous state flips. (2) *Native Temporal Encoding*: Models such as Qwen2.5-VL/Qwen3-VL/Qwen3.5 (Bai et al., 2025a;b; Qwen Team, 2026), InternVL3.5 (Wang et al., 2025b), Eagle2.5 (Chen et al., 2025), and Molmo2 (Clark et al., 2026) employ 3D rotary position embeddings (3D-RoPE) or explicit time-interval tokens to ground visual features in absolute timeline coordinates. Our diagnostics report overconfidence under observation gaps and lower accuracy on older-evidence queries among models using these representations.

D Additional Experimental Results

D.1 Complete six-cell decompositions

Figure 7 visualizes joint correctness patterns on Cook. Table 12 provides the full six-cell decomposition for all 14 models, alongside lowest-cell joint accuracy, status overconfidence, and STORM-BR-ATTR. Tables 13 and 14 report corresponding six-cell breakdowns on Health and Sports. Table 15 details online versus full-video offline performance.

Table 12: **Complete six-cell decomposition on Cook under the online protocol (RQ3).** Each cell reports joint answer–status accuracy (%). K/U denote Known/Uncertain ground truth; L/M/H denote change-intensity bins (1–3, 4–6, and 7–10). Worst indicates the minimum cell; OC is the overconfidence rate on Uncertain items. BR-Attr is the unsmoothed harmonic score with item-level cause-F1 modulation on Uncertain items (Eq. (4)).

	Low Change		Medium Change		High Change		Summary		
Model	**K**	**U**	**K**	**U**	**K**	**U**	**Worst**	**OC**	**BR-Attr**
Qwen2.5-VL-3B (2025a)	57.79	10.53	53.12	10.43	49.72	12.74	10.43	66.41	9.85
Qwen2.5-VL-7B (2025a)	57.08	21.05	52.67	23.91	48.79	21.70	21.05	38.42	16.28
Qwen3-VL-4B (2025b)	58.64	17.11	48.92	20.43	43.18	16.51	16.51	20.66	13.14
Qwen3-VL-8B (2025b)	70.25	18.42	65.83	5.22	63.93	8.02	5.22	41.12	8.37
Qwen3.5-4B (2026)	75.78	25.00	70.49	25.22	63.93	17.92	17.92	41.70	21.56
Qwen3.5-9B (2026)	82.86	22.37	76.39	22.17	67.10	18.87	18.87	45.95	19.15
InternVL3.5-8B (2025b)	78.19	6.58	75.14	3.91	74.39	3.30	3.30	80.50	4.34
Molmo2-8B (2026)	83.71	18.42	80.25	24.78	76.26	22.17	18.42	58.49	20.49
MiniCPM-V 4.5 (2025)	81.87	14.47	75.03	12.61	73.46	10.85	10.85	72.59	12.36
InternVideo2.5-8B (2025c)	35.69	11.84	39.05	6.96	36.64	6.13	6.13	27.99	5.40
Eagle2.5-8B (2025)	77.20	14.47	75.82	12.61	71.03	13.68	12.61	65.06	13.22
VideoLLaMA3-7B (2025)	88.67	7.89	81.73	5.22	76.26	4.25	4.25	86.87	5.06
LLaVA-NeXT-Video-7B (2024)	50.57	0.00	46.54	0.00	44.86	0.00	0.00	100.00	0.00
GLM-4.1V-9B-Thinking (2025)	24.50	27.63	17.82	28.26	13.46	25.00	13.46	5.79	15.83

Table 13: **Transfer decomposition: complete six-cell reliability on Health under the online protocol (RQ3).** Each cell reports joint answer–status accuracy (%). K/U denote Known/Uncertain ground truth and L/M/H denote change-intensity bins (1–3, 4–6, and 7–10). Occupancy is L = 260 Known / 29 Uncertain, M = 284 / 86, and H = 98 / 43. Worst indicates the minimum cell; OC is the overconfidence rate on Uncertain items. BR is the Laplace-smoothed six-cell STORM-BR.

	Low Change		Medium Change		High Change		Summary		
Model	**K**	**U**	**K**	**U**	**K**	**U**	**Worst**	**OC**	**BR**
Qwen2.5-VL-3B (2025a)	9.62	24.14	8.45	22.09	10.20	16.28	8.45	10.76	13.48
Qwen2.5-VL-7B (2025a)	7.31	41.38	3.17	32.56	3.06	18.60	3.06	1.27	7.78
Qwen3-VL-4B (2025b)	54.62	31.03	35.92	20.93	34.69	18.60	18.60	23.42	29.71
Qwen3-VL-8B (2025b)	56.92	17.24	49.30	9.30	51.02	11.63	9.30	31.65	21.28
Qwen3.5-4B (2026)	56.15	27.59	50.00	33.72	53.06	16.28	16.28	32.91	33.95
Qwen3.5-9B (2026)	58.46	24.14	48.94	24.42	52.04	20.93	20.93	25.32	33.23
InternVL3.5-8B (2025b)	57.31	13.79	53.87	5.81	65.31	9.30	5.81	68.99	17.14
Molmo2-8B (2026)	60.77	37.93	57.75	19.77	63.27	13.95	13.95	58.23	31.80
MiniCPM-V 4.5 (2025)	56.15	17.24	49.65	15.12	55.10	4.65	4.65	45.57	18.71
InternVideo2.5-8B (2025c)	34.23	10.34	36.62	9.30	30.61	2.33	2.33	25.32	12.27
Eagle2.5-8B (2025)	54.62	6.90	52.82	12.79	58.16	4.65	4.65	48.10	15.74
VideoLLaMA3-7B (2025)	62.69	17.24	54.23	2.33	60.20	2.33	2.33	86.71	9.66
LLaVA-NeXT-Video-7B (2024)	31.92	3.45	33.45	0.00	41.84	0.00	0.00	95.57	3.82
GLM-4.1V-9B-Thinking (2025)	26.92	41.38	12.68	27.91	15.31	23.26	12.68	5.06	21.69

D.2 CROSS-DOMAIN RANK STABILITY AND SIM RESULTS

Table 16 reports pairwise rank correlations (τ_b) and Kendall's W across real domains. Table 17 analyzes status gaps across the five real domains. Table 18 details behavioral shifts transferring from Cook to STORM-Sim, and Table 19 provides the 14-model simulation evaluation matrix.

Table 14: **Transfer decomposition: complete six-cell reliability on Sports under the online protocol (RQ3).** Each cell reports joint answer–status accuracy (%). K/U denote Known/Uncertain ground truth and L/M/H denote change-intensity bins (1–3, 4–6, and 7–10). Occupancy is L = 356 Known / 50 Uncertain, M = 221 / 78, and H = 63 / 32. Worst indicates the minimum cell; OC is the overconfidence rate on Uncertain items. BR is the Laplace-smoothed six-cell STORM-BR.

	Low Change		Medium Change		High Change		Summary		
Model	**K**	**U**	**K**	**U**	**K**	**U**	**Worst**	**OC**	**BR**
Qwen2.5-VL-3B (2025a)	6.18	22.00	4.52	12.82	9.52	15.62	4.52	13.12	9.62
Qwen2.5-VL-7B (2025a)	1.97	24.00	2.26	19.23	1.59	28.12	1.59	3.75	4.73
Qwen3-VL-4B (2025b)	38.20	10.00	28.96	14.10	31.75	21.88	10.00	23.75	20.89
Qwen3-VL-8B (2025b)	39.33	8.00	37.10	8.97	30.16	9.38	8.00	36.25	16.05
Qwen3.5-4B (2026)	34.55	10.00	34.84	8.97	31.75	15.62	8.97	34.38	18.09
Qwen3.5-9B (2026)	41.57	18.00	24.89	15.38	28.57	25.00	15.38	17.50	24.06
InternVL3.5-8B (2025b)	59.27	6.00	57.92	3.85	50.79	0.00	0.00	86.88	8.29
Molmo2-8B (2026)	54.49	16.00	49.77	23.08	52.38	28.12	16.00	56.88	31.34
MiniCPM-V 4.5 (2025)	49.72	6.00	50.68	8.97	42.86	15.62	6.00	68.75	17.16
InternVideo2.5-8B (2025c)	25.28	12.00	28.05	5.13	23.81	12.50	5.13	28.75	14.37
Eagle2.5-8B (2025)	40.73	6.00	46.61	7.69	42.86	3.12	3.12	58.13	12.41
VideoLLaMA3-7B (2025)	65.17	8.00	53.39	6.41	49.21	9.38	6.41	87.50	15.93
LLaVA-NeXT-Video-7B (2024)	31.46	0.00	37.56	0.00	33.33	0.00	0.00	97.50	3.43
GLM-4.1V-9B-Thinking (2025)	12.64	20.00	4.52	16.67	6.35	21.88	4.52	5.62	10.76

Table 15: **Full-video versus online STORM-BR on Cook (RQ2).** Both columns report the six-cell STORM-BR metric. Δ denotes full-video minus online performance in percentage points.

Model	**Full**	**Online**	**Δ**	**Model**	**Full**	**Online**	**Δ**
Qwen2.5-VL-3B	18.52	19.22	-0.71	MiniCPM-V 4.5	21.87	22.17	-0.30
Qwen2.5-VL-7B	31.32	31.60	-0.28	InternVideo2.5	13.33	13.39	-0.06
Qwen3-VL-4B	26.69	26.77	-0.08	Eagle2.5-8B	21.80	23.63	-1.83
Qwen3-VL-8B	15.99	15.21	+0.78	VideoLLaMA3	10.02	11.10	-1.08
Qwen3.5-4B	32.01	34.04	-2.03	LLaVA-NeXT	1.13	1.13	0.00
Qwen3.5-9B	29.33	33.29	-3.96	GLM-4.1V-Think	19.63	21.40	-1.77
InternVL3.5	8.64	8.97	-0.33	–	–	–	–
Molmo2-8B	36.83	34.45	+2.38	–	–	–	–

D.3 GLOBAL DIAGNOSTICS AND COOK ATTRIBUTION RESULTS

Table 20 reports domain-macro diagnostics, and Figure 8 summarizes question types and evidence-age slices. Table 21 provides model breakdowns across the six question types and recency slices. Table 22 details median performance and support counts across the six uncertainty root causes.

D.4 SMOOTHING AND PARTITION SENSITIVITY

Table 23 evaluates STORM-BR stability with and without Laplace smoothing on frozen model predictions. Smoothing preserves the top model across the six audited subsets, with rank correlations of 0.91–1.00, while assigning positive scores to cells lacking jointly correct predictions.

Table 24 evaluates ranking robustness under coarser binary aggregation and leave-one-cell-out sweeps. Collapsing the intensity bins preserves the top-ranked model in each subset; the six-cell versus collapsed rank correlation ranges from 0.52 on THOR to 0.98 on Cook and Sports.

Table 25 evaluates sensitivity to the intensity partition by sweeping 15 valid alternative cutpoint pairs around the canonical $(3, 6)$ boundary on identical frozen predictions. Across all alternatives, the canonical leader remains within the top-3 across all audited subsets, and rank correlation with the canonical metric satisfies median $\tau \geq 0.93$ across four of the five original complete domains

Table 16: **What transfers across domains: model-rank agreement on STORM-Real (RQ1).** Kendall's τ_b is computed across all 14 models for each domain pair. Known and Uncertain are status-conditioned joint accuracies. Mean τ_b summarises the 10 pairs; Kendall's W measures joint agreement across all five domains.

Domain pair	Accuracy	Known joint	Uncertain joint	STORM-BR
Cook–Bike	0.71	0.54	0.80	0.34
Cook–Health	0.84	0.67	0.76	0.54
Cook–Music	0.54	0.64	0.79	0.54
Cook–Sports	0.85	0.69	0.71	0.52
Bike–Health	0.60	0.69	0.84	0.36
Bike–Music	0.42	0.64	0.76	0.54
Bike–Sports	0.87	0.86	0.68	0.65
Health–Music	0.57	0.80	0.76	0.65
Health–Sports	0.73	0.84	0.73	0.58
Music–Sports	0.47	0.79	0.87	0.80
Mean pairwise τ_b	0.66	0.72	0.77	0.55
Five-domain Kendall's W	0.83	0.87	0.92	0.75

Table 17: **Which side is the bottleneck: status-conditioned reliability balance across STORM-Real domains (RQ1, RQ3).** Each value is Known minus Uncertain joint accuracy in percentage points. Positive values indicate that Uncertain performance is limiting; negative values indicate that Known performance is limiting.

Model	Cook	Bike	Health	Music	Sports
Qwen2.5-VL-3B	+42.4	-9.2	-11.7	-33.3	-10.3
Qwen2.5-VL-7B	+30.6	-26.0	-25.6	-45.8	-20.5
Qwen3-VL-4B	+32.4	+25.2	+21.2	-3.1	+20.0
Qwen3-VL-8B	+58.5	+48.4	+41.3	+22.9	+28.9
Qwen3.5-4B	+48.4	-24.2	+25.1	+1.0	+23.8
Qwen3.5-9B	+55.4	+1.2	+29.9	+16.7	+16.4
InternVL3.5-8B	+71.9	+73.5	+48.8	+49.0	+54.2
Molmo2-8B	+57.6	+51.1	+38.3	+11.5	+30.8
MiniCPM-V 4.5	+64.7	+48.7	+40.5	+35.4	+40.0
InternVideo2.5-8B	+30.0	+24.8	+27.1	-9.4	+17.3
Eagle2.5-8B	+61.8	+46.0	+44.9	+22.9	+36.7
VideoLLaMA3-7B	+77.4	+69.7	+53.5	+39.6	+52.0
LLaVA-NeXT-Video-7B-DPO	+47.5	+43.4	+33.5	+28.1	+33.8
GLM-4.1V-9B-Thinking	-7.9	-15.8	-10.3	-31.2	-9.5

($\tau = 0.76$ on THOR). The canonical top-3 set is fully preserved on Bike, Sports, and in 14 of 15 Cook configurations, while the corresponding top-3 sets are preserved in 6 of 15 Health partitions and 3 of 15 THOR partitions. VHome admits 12 of the same 15 alternatives with six occupied cells: its median τ is 0.78, its leader remains in the top three in all 12, and its full top-three set changes in every alternative.

D.5 COOK INPUT CONTROLS

Figure 8d summarizes absolute scores for the 42 input configurations on Cook. Table 26 reports detailed model-level scores across all controls. Table 27 reports online, full-video, text-only, shuffled-frame, and 16-frame recent-history results for two representative models.

Table 18: **Cook-to-Sim shifts in model behaviour (RQ1).** Deltas are the indicated simulation subset minus Cook in percentage points. Known/Uncertain denote status-conditioned joint accuracy; OC and UC denote overconfidence and underconfidence rates.

Model	△Acc.	△Known	△Uncertain	△OC	△UC	△BR
THOR minus Cook						
Qwen2.5-VL-3B	-8.67	-27.69	+4.97	-59.14	+48.15	-8.92
Qwen2.5-VL-7B	-13.39	-38.06	-4.41	-38.42	+58.32	-21.43
Qwen3-VL-4B	-14.36	-1.32	+8.93	+2.98	-14.78	-7.50
Qwen3-VL-8B	-20.66	-21.52	+4.43	-17.48	+4.78	-3.98
Qwen3.5-4B	-18.26	-54.27	+8.71	-41.70	+64.09	-18.62
Qwen3.5-9B	-19.73	-49.26	+28.24	-45.95	+55.36	-7.40
InternVL3.5-8B	-19.89	-26.58	+1.40	-16.87	+1.33	+5.03
Molmo2-8B	-19.99	-22.44	+13.58	-23.95	-2.25	-0.30
MiniCPM-V 4.5	-22.80	-25.89	-1.25	-34.41	+0.95	-0.47
InternVideo2.5-8B	-11.86	-5.49	+10.85	+2.92	+15.32	-2.75
Eagle2.5-8B	-24.45	-31.40	-4.23	+7.67	+10.53	-8.19
VideoLLaMA3-7B	-23.66	-28.78	-3.39	+5.85	-0.66	-1.78
LLaVA-NeXT-Video-7B-DPO	-1.67	-3.37	+0.00	-14.55	+4.88	+7.10
GLM-4.1V-9B-Thinking	-15.93	-1.39	+4.08	-5.79	+0.53	-10.06
VHome minus Cook						
Qwen2.5-VL-3B	-15.74	-23.86	-5.73	-17.35	+22.43	-5.98
Qwen2.5-VL-7B	-24.16	-37.55	+5.72	-6.34	+45.95	-10.10
Qwen3-VL-4B	-22.92	-29.19	+0.53	-13.11	+18.89	-7.81
Qwen3-VL-8B	-22.25	-41.93	+4.91	-24.14	+40.13	+3.82
Qwen3.5-4B	-23.60	-59.20	+11.76	-41.70	+68.12	-17.39
Qwen3.5-9B	-29.16	-66.50	+7.45	-45.95	+73.34	-18.19
InternVL3.5-8B	-20.12	-44.74	-0.28	-57.86	+37.73	+3.53
Molmo2-8B	-29.43	-49.59	+33.82	-28.31	+28.70	+5.89
MiniCPM-V 4.5	-26.71	-43.58	-0.84	-8.44	+30.26	-1.84
InternVideo2.5-8B	-11.91	-25.09	+9.65	-16.67	+34.39	+2.07
Eagle2.5-8B	-26.32	-56.08	+1.77	-53.74	+57.58	-4.18
VideoLLaMA3-7B	-24.72	-29.92	-5.21	+1.81	+4.87	-1.79
LLaVA-NeXT-Video-7B-DPO	-13.85	-28.89	+0.00	-58.49	+41.35	+6.97
GLM-4.1V-9B-Thinking	-11.93	-11.77	+46.75	-5.79	+12.07	-7.37

Table 19: **Controlled stress test: complete 14-model reliability on STORM-Sim (RQ1, RQ2).** THOR contains 300 items over 28 AI2-THOR episodes; VHome adds 290 items over 29 VirtualHome episodes. Full-video and Online report task-choice accuracy with the complete sequence and the query-time prefix, respectively; both use the original query-time target labels. KCI, Epi., and Storm-BR use online predictions. KCI is Known Change-Intensity; Epi. is the Known/Uncertain joint-accuracy macro-average. STORM-BR is the Laplace-smoothed harmonic mean over the six cells, with 95% episode-clustered bootstrap intervals (1,000 resamples). In THOR, Low-change Uncertain has n=8; that cell is a transfer diagnostic. Best and second-best in each column are bold and underlined. All values are percentages.

Model	Full-video	Online	KCI	Epi.	STORM-BR
THOR					
Qwen2.5-VL-3B (2025a)	40.33	45.00	45.35	21.24	10.30 [7.7–11.9]
Qwen2.5-VL-7B (2025a)	44.00	47.67	44.88	16.64	10.17 [7.4–11.6]
Qwen3-VL-4B (2025b)	**48.00**	51.67	<u>53.67</u>	<u>38.33</u>	19.27 [10.8–27.7]
Qwen3-VL-8B (2025b)	41.00	44.00	48.21	29.02	11.23 [8.0–12.9]
Qwen3.5-4B (2026)	**48.00**	55.00	51.68	23.62	15.42 [8.3–20.0]
Qwen3.5-9B (2026)	<u>47.33</u>	**56.67**	**54.03**	38.01	<u>25.89</u> [12.6–33.2]
InternVL3.5-8B (2025b)	40.00	43.33	42.05	27.42	14.00 [8.4–17.2]
Molmo2-8B (2026)	46.00	<u>56.33</u>	50.56	**47.16**	**34.15** [21.6–42.4]
MiniCPM-V 4.5 (2025)	42.33	46.33	46.27	30.96	21.70 [11.5–26.2]
InternVideo2.5-8B (2025c)	41.33	48.67	50.83	25.01	10.64 [7.7–12.8]
Eagle2.5-8B (2025)	40.67	45.67	49.21	26.38	15.44 [7.6–21.4]
VideoLLaMA3-7B (2025)	41.00	49.33	48.34	27.85	9.32 [7.6–10.7]
LLaVA-NeXT-Video-7B-DPO (2024)	35.67	36.67	36.98	22.04	8.23 [6.5–9.4]
GLM-4.1V-9B-Thinking (2025)	41.00	46.00	45.37	24.23	11.34 [8.8–13.0]
VHome					
Qwen2.5-VL-3B (2025a)	38.97	37.93	49.15	17.81	13.25 [8.8–15.6]
Qwen2.5-VL-7B (2025a)	35.52	36.90	38.41	21.96	<u>21.50</u> [14.0–26.5]
Qwen3-VL-4B (2025b)	44.48	43.10	49.26	20.19	18.96 [11.0–24.9]
Qwen3-VL-8B (2025b)	42.41	42.41	49.44	19.05	19.04 [10.2–25.4]
Qwen3.5-4B (2026)	**52.07**	<u>49.66</u>	53.54	22.68	16.65 [8.1–20.7]
Qwen3.5-9B (2026)	<u>46.55</u>	47.24	52.45	19.00	15.10 [9.5–18.1]
InternVL3.5-8B (2025b)	40.69	43.10	53.10	17.50	12.50 [8.3–17.4]
Molmo2-8B (2026)	37.59	46.90	42.01	**43.70**	**40.35** [31.8–44.9]
MiniCPM-V 4.5 (2025)	42.07	42.41	51.71	22.33	20.33 [10.9–27.5]
InternVideo2.5-8B (2025c)	44.14	48.62	<u>59.01</u>	14.61	15.45 [7.7–19.9]
Eagle2.5-8B (2025)	42.76	43.79	53.22	17.04	19.45 [11.3–22.8]
VideoLLaMA3-7B (2025)	41.38	48.28	**60.60**	26.37	9.31 [8.8–9.9]
LLaVA-NeXT-Video-7B-DPO (2024)	24.14	24.48	31.23	9.28	8.10 [7.4–8.6]
GLM-4.1V-9B-Thinking (2025)	42.41	**50.00**	47.60	<u>40.38</u>	14.03 [5.8–19.9]

Table 20: Domain-macro diagnostics from frozen online predictions. Known (K) and Uncertain (U) joint accuracy and history-aggregation task accuracy (Hist.) average seven subsets equally. Recent–Old is the joint-accuracy difference between the 0–2s and $>$ 5s slices, averaged over the five real domains and VHome; THOR has no old-evidence items. All values are percentages or percentage-point differences.

Model	K joint	U joint	Hist.	Recent–Old
Q2.5-VL-3B	19.55	17.85	32.66	+0.94
Q2.5-VL-7B	13.80	28.52	31.39	-3.53
Q3-VL-4B	40.08	22.88	42.55	-3.69
Q3-VL-8B	46.30	11.40	34.82	+5.47
Q3.5-4B	33.30	28.02	34.10	+2.37
Q3.5-9B	40.05	28.80	35.61	+6.88
IVL3.5-8B	57.18	4.49	41.53	+3.86
Molmo2-8B	57.83	31.39	46.55	+5.09
MCPM-V 4.5	54.59	12.95	42.57	+4.34
IVideo2.5-8B	29.02	14.91	36.53	-0.05
Eagle2.5-8B	48.21	12.40	40.97	-1.93
VLLM3-7B	62.28	5.55	40.58	+8.64
LLaVA-7B-DPO	35.64	0.09	23.80	+7.00
GLM-4.1V	13.74	35.81	35.87	+2.95

Table 21: **What temporal tracking requires: semantic-type and evidence-age results on Cook (RQ4).** The six type columns report online accuracy (%). Recent and Old report joint answer–status accuracy for evidence ending 0–2 s and more than 5 s before the query. All values cover the same 14-model panel visualized in Figure 8a,b.

Model	Current	Fact	History	Change	Track	Temporal	Recent	Old
Qwen2.5-VL-3B	53.81	64.80	29.19	66.74	63.74	44.09	46.77	37.85
Qwen2.5-VL-7B	58.74	70.63	35.97	72.44	74.10	54.77	48.48	38.49
Qwen3-VL-4B	65.92	74.36	45.48	72.44	76.80	61.36	44.35	46.06
Qwen3-VL-8B	63.68	73.43	42.99	73.35	71.62	63.18	57.35	43.53
Qwen3.5-4B	75.11	84.62	47.74	82.69	81.98	67.73	62.24	54.89
Qwen3.5-9B	78.03	85.31	55.66	83.83	84.01	71.82	66.50	58.99
InternVL3.5-8B	61.88	72.73	43.21	71.98	68.47	61.36	64.08	48.26
Molmo2-8B	79.82	80.19	56.79	82.92	83.78	74.55	71.48	54.57
MiniCPM-V 4.5	68.16	75.99	51.81	74.94	78.60	65.45	66.19	52.05
InternVideo2.5-8B	60.99	64.10	41.40	69.70	66.67	60.45	31.66	30.91
Eagle2.5-8B	67.94	78.32	47.51	77.22	81.08	68.86	65.02	49.53
VideoLLaMA3-7B	76.01	85.55	48.19	78.82	84.46	65.23	70.31	49.84
LLaVA-NeXT-Video-7B-DPO	43.05	33.33	23.30	47.61	52.03	30.45	40.76	23.34
GLM-4.1V-9B-Thinking	65.70	66.20	39.82	74.26	72.75	52.95	21.26	17.35

Table 22: **Where uncertainty attribution fails: source-conditioned results on Cook (RQ3).** Values are medians over 14 models. Joint is answer–status correctness conditioned on each ground-truth source; Attribution is the source-specific F1. Sources are multi-label and may overlap.

Source	n	Joint	Attribution F1
Missing observation	405	11.23	54.14
Partial observation	187	13.37	1.04
Ambiguous evidence	142	16.55	1.40
Low visual quality	71	34.51	24.07
Multiple candidates	26	17.31	3.12
Ambiguous attribute	15	40.00	0.00

Table 23: **Metric stability: raw versus add-one-smoothed STORM-BR (RQ2).** Both scores are recomputed from identical frozen prediction files. Kendall τ compares the 14-model raw and smoothed rankings, with tied pairs dropped. $|\Delta|_{\max}$ is the largest model-wise score change in percentage points; a raw zero occurs when at least one joint cell has no correct prediction. Smoothing does not change the leader on any listed domain, while making rare-cell scores finite and comparable.

Domain	τ rank	Raw top	Smooth top	$\lvert\Delta\rvert_{\max}$ (pp)	Zero / 14
Cook	1.00	Molmo2-8B	Molmo2-8B	1.13	1
Bike	0.95	Molmo2-8B	Molmo2-8B	8.72	3
Health	0.96	Qwen3.5-4B	Qwen3.5-4B	3.98	1
Sports	0.93	Molmo2-8B	Molmo2-8B	8.29	2
THOR	0.97	Molmo2-8B	Molmo2-8B	14.00	8
VHome	0.91	Molmo2-8B	Molmo2-8B	13.25	4

Table 24: **Aggregation robustness of STORM-BR on sparse cells (RQ2).** The collapsed variant pools the three change-intensity bins into a single Known/Uncertain pair; leave-one-cell-out (LOO) drops each of the six cells in turn. τ is the Kendall rank correlation between the 14-model six-cell and collapsed scores, $|\Delta|_{\max}$ is the largest model-wise score change, and the LOO column summarizes score sensitivity to single-cell deletion. Collapsing preserves the leader on every evaluated subset, while THOR changes scores most (up to 16.5 pp).

Domain	n	τ	$\lvert\Delta\rvert_{\max}$ **(pp)**	**LOO span (pp)**	**Leader**
Cook	2,640	0.98	1.17	6.0–10.5	Molmo2-8B
Bike	786	0.91	5.53	6.3–14.3	Molmo2-8B
Health	800	0.93	2.78	7.4–12.8	Qwen3.5-4B
Sports	800	0.98	2.22	5.1–8.5	Molmo2-8B
THOR	300	0.52	16.52	8.7–15.9	Molmo2-8B
Music	120	n/a	n/a	n/a	Molmo2-8B
VHome	290	0.87	5.80	2.3–13.5	Molmo2-8B

Table 25: **Sensitivity to alternative change-intensity partitions (RQ2).** Each of the 15 alternative partitions reuses the same 14 frozen online prediction files and changes only the two cutpoints that define the Low/Medium/High bins (canonical cuts $(3, 6)$). The alternatives are the nearest valid cutpoint pairs to $(3, 6)$; validity requires six non-empty intensity–answerability cells. VHome uses the 12 of these 15 alternatives that satisfy this condition; the three cuts starting at 1 leave its Low bin empty. τ is the Kendall rank correlation between the 14-model ranking and the canonical ranking. "Leader in top-3" counts partitions in which the canonical leader remains within the top three, "Top-3 set" counts partitions that preserve the canonical top-3 set, and $|\Delta|_{\max}$ is the largest change among the canonical top-3 scores in percentage points. The last column is the fraction of model×partition comparisons that preserve the sign of the canonical Known–Uncertain reliability gap. Music lacks high-intensity queries (B_3) and is excluded from six-cell partition audits.

Domain	τ **median [min]**	**Leader in top-3**	**Top-3 set**	$\lvert\Delta\rvert_{\max}$ **(pp)**	**Known–Unc. sign**
Cook	0.96 [0.91]	15/15	14/15	4.89	100.0%
Bike	0.93 [0.85]	15/15	15/15	9.63	98.1%
Health	0.96 [0.78]	15/15	6/15	7.76	100.0%
Sports	0.96 [0.89]	15/15	15/15	7.21	100.0%
THOR	0.76 [0.69]	15/15	3/15	9.36	89.0%
VHome	0.78 [0.67]	12/12	0/12	4.57	95.2%

Table 26: **Input configurations on Cook (RQ4).** All 14 checkpoints are evaluated under text-only, shuffled-frame, and recent-16-frame input. Acc. is task accuracy; KCI is accuracy averaged over intensity bins; Epi. is Known/Uncertain joint-accuracy macro-average; BR is add-one-smoothed six-cell STORM-BR.

Model	Control	Sampling / cap	Acc.	KCI	Epi.	BR
Qwen2.5-VL-3B	Question + choices	no video	35.08	34.88	19.75	15.70
Qwen2.5-VL-3B	Shuffled frames	1 FPS / all	39.89	39.58	12.68	8.34
Qwen2.5-VL-3B	Recent history	1 FPS / 16	42.77	42.64	15.35	11.51
Qwen2.5-VL-7B	Question + choices	no video	42.54	42.34	30.31	0.28
Qwen2.5-VL-7B	Shuffled frames	1 FPS / all	46.48	46.40	17.00	2.84
Qwen2.5-VL-7B	Recent history	1 FPS / 16	48.64	48.76	19.61	4.53
Qwen3-VL-4B	Question + choices	no video	37.27	36.96	15.29	1.32
Qwen3-VL-4B	Shuffled frames	1 FPS / all	46.06	46.32	22.83	22.12
Qwen3-VL-4B	Recent history	1 FPS / 16	48.86	49.12	26.06	23.92
Qwen3-VL-8B	Question + choices	no video	35.80	35.60	10.93	0.85
Qwen3-VL-8B	Shuffled frames	1 FPS / all	45.64	45.91	25.66	16.51
Qwen3-VL-8B	Recent history	1 FPS / 16	47.95	48.27	27.63	18.68
Qwen3.5-4B	Question + choices	no video	41.17	41.10	29.05	0.28
Qwen3.5-4B	Shuffled frames	1 FPS / all	55.87	55.92	38.39	36.54
Qwen3.5-4B	Recent history	1 FPS / 16	59.55	59.84	41.58	40.41
Qwen3.5-9B	Question + choices	no video	39.17	39.09	28.31	0.34
Qwen3.5-9B	Shuffled frames	1 FPS / all	57.54	57.55	36.28	33.56
Qwen3.5-9B	Recent history	1 FPS / 16	61.89	61.80	39.91	35.49
InternVL3.5-8B	Question + choices	no video	35.98	35.85	16.68	15.44
InternVL3.5-8B	Shuffled frames	1 FPS / all	46.55	46.65	30.31	10.16
InternVL3.5-8B	Recent history	1 FPS / 16	51.25	51.54	33.41	11.74
Molmo2-8B	Question + choices	no video	49.24	49.00	36.57	18.96
Molmo2-8B	Shuffled frames	1 FPS / all	54.85	54.76	37.06	26.57
Molmo2-8B	Recent history	1 FPS / 16	59.13	59.21	39.97	26.79
MiniCPM-V 4.5	Question + choices	no video	43.71	43.48	30.10	29.75
MiniCPM-V 4.5	Shuffled frames	1 FPS / all	54.20	54.44	36.75	29.49
MiniCPM-V 4.5	Recent history	1 FPS / 16	58.07	58.24	40.89	33.19
InternVideo2.5-8B	Question + choices	no video	37.69	37.45	13.15	0.89
InternVideo2.5-8B	Shuffled frames	1 FPS / all	43.67	43.52	18.87	15.25
InternVideo2.5-8B	Recent history	1 FPS / 16	46.02	45.93	19.49	15.62
Eagle2.5-8B	Question + choices	no video	42.08	41.62	23.01	2.55
Eagle2.5-8B	Shuffled frames	1 FPS / all	53.75	53.77	32.53	21.90
Eagle2.5-8B	Recent history	1 FPS / 16	57.95	57.93	38.19	31.69
VideoLLaMA3-7B	Question + choices	no video	42.50	42.38	25.10	25.57
VideoLLaMA3-7B	Shuffled frames	1 FPS / all	53.14	53.24	31.67	9.22
VideoLLaMA3-7B	Recent history	1 FPS / 16	59.28	59.36	36.43	15.15
LLaVA-NeXT-Video-7B-DPO	Question + choices	no video	25.91	26.00	15.76	1.12
LLaVA-NeXT-Video-7B-DPO	Shuffled frames	1 FPS / all	31.36	31.37	19.33	1.59
LLaVA-NeXT-Video-7B-DPO	Recent history	1 FPS / 16	31.48	31.49	19.51	1.44
GLM-4.1V-9B-Thinking	Question + choices	no video	42.69	42.17	22.97	0.63
GLM-4.1V-9B-Thinking	Shuffled frames	1 FPS / all	40.42	40.33	20.49	9.38
GLM-4.1V-9B-Thinking	Recent history	1 FPS / 16	42.77	42.75	23.07	12.78

Table 27: **Cook protocol and input-configuration results.** Performance of two representative models under official online, full-video offline, and input-control configurations. Acc. denotes task accuracy, and BR denotes balanced reliability (STORM-BR).

Model	Input / control	Sampling / cap	Acc.	BR
Qwen3.5-9B (2026)	Official online	1 FPS / all	76.40	33.29
	Full-video offline	1 FPS / all	75.04	29.33
	Question + choices	no video	39.17	0.34
	Bounded recent history	1 FPS / 16	61.89	35.49
	Shuffled visible frames	1 FPS / all	57.54	33.56
Molmo2-8B (2026)	Official online	1 FPS / all	76.33	34.45
	Full-video offline	1 FPS / all	76.40	36.83
	Question + choices	no video	49.24	18.96
	Bounded recent history	1 FPS / 16	59.13	26.79
	Shuffled visible frames	1 FPS / all	54.85	26.57

E LLM USAGE

In accordance with disclosure guidelines, we detail language model use across benchmark construction and paper preparation. In the real track, initial candidate QA proposals were synthesized using Doubao Seed 2.1 Pro via the Volcengine Ark API, and multimodal LLMs assisted in proposing plausible distractors. In the simulation track, VLMs assisted in refining template syntax during development (Appendix A.2.7), while final QA items were generated via deterministic rule engines paired with simulator execution logs (Appendix A.2.4). Real-world candidate items underwent three-co-author review of answers, options, evidence spans, epistemic labels, and uncertainty sources (Appendix A.7). Simulation items underwent the subset-specific structural and evidence checks in Appendix A.2, including targeted VHome RGB endpoint review for physical-state questions. LLMs assisted manuscript proofreading and LaTeX macro formatting. Reported metrics were computed strictly from raw predictions using deterministic evaluation scripts.