From Finding to Linking: Benchmarking and Advancing Cross-Long-Video Reasoning for Multimodal LLMs

Meng Luo **Zikang Zhou** **Shanqing Xu** **Shize Zhang** **Bobo Li***
Hao Fei **Mong-Li Lee** **Wynne Hsu**
National University of Singapore

Abstract

Reasoning across multiple long-form videos requires models to *find* sparse evidence within each video and *link* related entities, events, and narratives across streams. Existing long-video benchmarks mainly evaluate single-video understanding, while multi-video benchmarks typically use short clips. We introduce **CLoVR-Bench**, a comprehensive benchmark for Cross-Long-Video Reasoning with 2,000 expert-annotated QA pairs over 400 long-form videos. Its three-level taxonomy covers Comparative Analysis, Tracking and Retrieval, and Integrated Reasoning, spanning 14 tasks and 36 subtasks that diagnose both intra-video localization and inter-video alignment. Our evaluation of 13 representative MLLMs shows substantial gaps across this hierarchy, with performance degrading from localized comparison to long-range retrieval and integrated cross-video reasoning. We further propose **HOLMES**, a training-free framework that formulates cross-long-video reasoning as evidence-slot filling over a typed binding graph. HOLMES plans option-discriminating evidence slots, localizes evidence with risk-conditioned policies, records coverage certificates for failed searches, verifies cross-video bindings visually, audits constraints, and returns answers through hard graph-readout gates. Experiments show that HOLMES improves both open-source and closed-source backbones, outperforms adapted long-video baselines, and yields the largest gains on tasks requiring long-range evidence search and cross-video binding. Together, CLoVR-Bench and HOLMES provide a diagnostic testbed and an interpretable baseline for future cross-long-video reasoning. CLoVR-Bench is publicly available at `https://huggingface.co/datasets/CLoVR-Bench/CLoVR-Bench`.

1 Introduction

Real-world video understanding rarely comes as a single short clip. Evidence is often scattered across long videos captured from different cameras, viewpoints, or episodes, and an answer is reliable only when the relevant moments are first found and then connected across streams [36, 35]. Applications such as video archive review and egocentric-memory synthesis therefore require more than isolated video comprehension: models must reason over entities, events, and narratives distributed across multiple long-form videos [30, 29, 18]. We study this setting as *cross-long-video reasoning*.

Cross-long-video reasoning poses two coupled challenges. First, *evidence finding* is difficult because each source video may last tens of minutes, while the decisive evidence can occupy only a few seconds. Uniform sampling and similarity retrieval [15, 31, 10] can easily miss visually subtle or temporally sparse events. Second, *evidence linking* is difficult because relevant people, objects, places, and narratives may recur across videos under different viewpoints, temporal orders, and appearances. Reliable reasoning therefore requires models not only to locate sparse evidence within each long video, but also to verify how pieces of evidence correspond across videos.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing benchmarks cover only adjacent parts of this regime. Long-video benchmarks [37, 41] stress temporal context, but primarily evaluate one video at a time. Multi-video benchmarks [24, 13] require cross-video comparison, but typically use short clips in which relevant evidence is dense and easy to expose. As a result, existing evaluations provide limited coverage of the joint regime required for realistic cross-long-video reasoning: finding sparse evidence in long streams and linking it across videos. To fill this gap, we introduce **CLoVR-Bench**, a comprehensive benchmark for **C**ross-**Lo**ng-**V**ideo **R**easoning with 2,000 expert-annotated QA pairs grounded in 400 real long-form videos, each at least 30 minutes long. CLoVR-Bench organizes questions into a three-level taxonomy: Comparative Analysis, Tracking and Retrieval, and Integrated Reasoning (Figure 1). These levels span 14 tasks and 36 subtasks annotated with temporal evidence intervals and cross-video links.

Our evaluation of 13 representative MLLMs reveals substantial gaps across this hierarchy. Accuracy drops from localized comparative analysis to tracking and retrieval, and further to integrated cross-video reasoning. The largest gaps between models and humans occur on Integrated Reasoning subtasks that require psychological and social reasoning, event reconstruction, and cross-video consistency checking. These results suggest that cross-long-video reasoning is not merely a long-context scaling problem. Models need mechanisms that can selectively search for sparse evidence, preserve uncertainty when evidence is not found, and explicitly verify bindings across videos.

To this end, we propose **HOLMES** (**H**ypothesis-driven **O**bservation, **L**inking, and **M**ulti-source **E**vidence **S**ynthesis), a training-free framework for structured evidence acquisition and visual binding. Instead of prompting an MLLM to answer from a large visual context, HOLMES builds an auditable evidence dossier: what evidence must be found, where it was found, how evidence across videos is linked, and whether the resulting graph satisfies the answer constraints. HOLMES plans evidence slots that discriminate among candidate answers without committing to one, localizes each slot with a risk-conditioned policy, records coverage certificates for failed searches, verifies cross-video bindings visually, audits constraints, and returns answers through hard graph-readout gates.

Experiments show that HOLMES consistently improves both open-source and closed-source MLLM backbones on CLoVR-Bench. The improvements are most pronounced on tasks that require long-range evidence search and cross-video binding, suggesting that performance depends not only on visual perception, but also on how evidence is found, linked, and verified. Comparisons with adapted long-video baselines further show that the proposed inference structure transfers across backbones and to public long-video benchmarks.

In summary, our main contributions are:

- We introduce **CLoVR-Bench**, a comprehensive benchmark for Cross-Long-Video Reasoning featuring 2,000 expert-annotated QA pairs over 400 real long-form videos, each at least 30 minutes long. It combines a hierarchical cognitive taxonomy with evidence-grounded annotations, including temporal evidence intervals and cross-video links.
- We systematically evaluate 13 MLLMs and adapted long-video baselines, revealing substantial performance gaps as tasks progress from perception-oriented comparison to long-range retrieval and integrated cross-video reasoning.
- We propose **HOLMES**, a training-free framework that turns cross-long-video QA from monolithic prompting into evidence-slot planning, risk-conditioned localization, typed visual binding, and hard-gated graph readout over an auditable evidence dossier.

2 Related Work

Video MLLMs. Video MLLMs extend language-model reasoning to videos by coupling visual encoders with language backbones and designing mechanisms for temporal aggregation [38, 33, 19]. As video duration increases, models must balance frame sampling density, visual-token compression, temporal encoding, and attention or memory efficiency. Recent systems improve long-video processing through dynamic resolution, temporal position modeling, visual-token reduction, and long-context language backbones. For example, Qwen2.5-VL [3] introduces dynamic-resolution processing and absolute time encoding for hour-scale video comprehension with fine-grained event localization. Despite this progress, current MLLMs remain brittle when evidence is sparse, temporally distant, and distributed across multiple long videos, where robust cross-video integration is required.

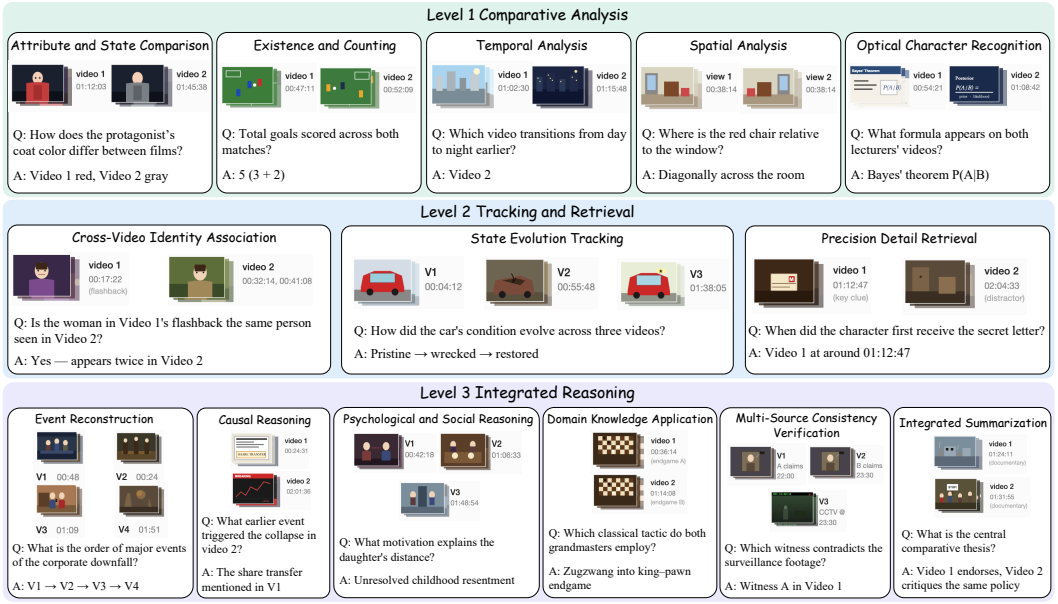

Figure 1: Overview of CLoVR-Bench. It evaluates cross-long-video reasoning across three hierarchical levels and 14 distinct tasks.

Video Understanding Benchmarks. Existing benchmarks for video understanding and video QA largely follow a single-video paradigm [20]. Early efforts such as MVBench [14] cover diverse temporal reasoning skills with short clips, while recent long-form benchmarks scale the temporal horizon to minutes or hours. LongVideoBench [37] introduces referring reasoning over interleaved video-language inputs, and MLVU [41] targets multi-task evaluation across diverse long-video genres. However, these long-video benchmarks still operate primarily on single-video inputs and therefore do not test cross-video evidence aggregation, comparison, and alignment. In parallel, multi-video benchmarks have begun to evaluate reasoning across videos. MVU-Eval [24], CVBench [43], and CrossVid [13] expand multi-video understanding, but their videos are much shorter than hour-scale long-form footage, leaving sparse-evidence cross-video reasoning underexplored. CLoVR-Bench complements these efforts by jointly targeting multiple videos and long-form temporal contexts, emphasizing evidence selection under noise, cross-video temporal alignment, and integrative reasoning over sparse information distributed across extended streams. Table 1 summarizes the key differences. CVBench and CrossVid also use hierarchical task organization; CLoVR-Bench specifically separates localized perception, long-range evidence finding, and cross-video integration, with evidence annotations that support the diagnostics in Appendix K.2.

Long Video Understanding and Reasoning. A growing line of work improves long video QA at inference time through uniform or adaptive sampling, keyframe selection, retrieval-based context construction, summarization, and agentic decomposition [45, 11, 28, 27]. These methods reduce the visual context presented to the MLLM and can improve single-video reasoning when relevant evidence can be retrieved or summarized reliably. However, they typically leave evidence aggregation implicit in a prompt, summary, or selected frame set, and are not designed to explicitly separate evidence finding from cross-video binding. In contrast, HOLMES separates finding, linking, and readout, making cross-video evidence aggregation inspectable.

3 CLoVR-Bench: Benchmarking Cross-Long-Video Reasoning

To systematically evaluate cross-long-video understanding beyond isolated task performance, CLoVR-Bench introduces a hierarchical taxonomy that decomposes this capability into progressively harder stages: seeing clearly, finding accurately, and reasoning coherently. Under this framework, CLoVR-Bench instantiates 14 task types and 36 subtasks. ❶ **Comparative Analysis** tests localized perception through Attribute and State Comparison (ASC), Existence and Counting (EC), Temporal Analysis (TA), Spatial Analysis (SA), and Optical Character Recognition (OCR). ❷ **Tracking and Retrieval** tests long-range evidence search through Cross-Video Identity Association (CIA), State Evolution

Table 1: Comparison of representative video benchmarks. Parentheses in the task column give subtask counts. Multi-video and Multi-view indicate whether a benchmark requires reasoning across videos and viewpoints, respectively.

Benchmark	# Videos	# QA	Avg. len. (s)	# Tasks	Multi-video	Multi-view
MVBench [14]	3,641	4,000	16	20	✗	✗
LongVideoBench [37]	3,763	6,678	473	17	✗	✗
Video-MME [8]	900	2,700	1,017	30	✗	✗
MLVU [41]	1,730	3,102	930	9	✗	✗
All-Angles Bench [39]	90	2,132	/	6	✗	✓
MVU-Eval [24]	4,959	1,824	<60	8	✓	✓
CVBench [43]	1,315	1,000	106.6	3 (15)	✓	✓
CrossVid [13]	5,331	9,015	215	10	✓	✓
CLoVR-Bench (Ours)	**400**	**2,000**	**3,018**	**14 (36)**	✓	✓

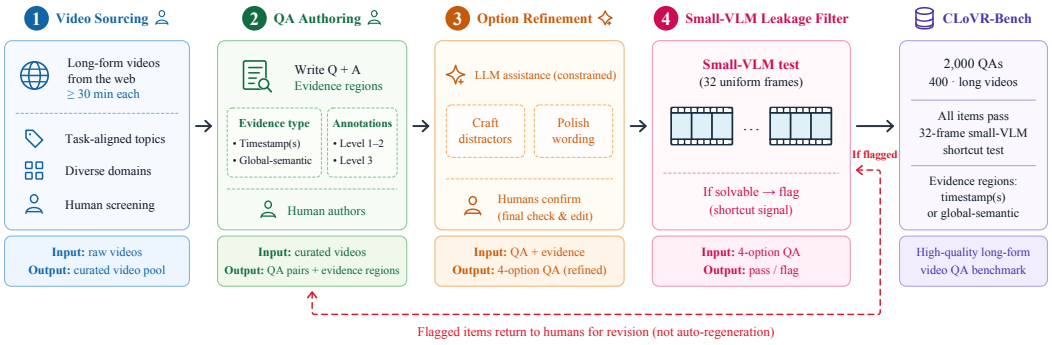

Figure 2: Pipeline of CLoVR-Bench data construction. The process consists of four main stages.

Tracking (SET), and Precision Detail Retrieval (PDR). ❸ **Integrated Reasoning** tests synthesis through Event Reconstruction (ER), Causal Reasoning (CR), Psychological and Social Reasoning (PSR) [17, 16], Domain Knowledge Application (DKA), Multi-Source Consistency Verification (MSCV), and Integrated Summarization (IS). Detailed definitions of all tasks are provided in Appendix A, and Figure 3 reports per-subtask item counts and group statistics.

We construct CLoVR-Bench through a human-led pipeline with targeted assistance from LLMs and VLMs. Expert annotators first curate long-form videos from publicly available web platforms and form source groups that match our subtask definitions, including spatiotemporal patterns, entity correspondences, and causal structures. This manual grouping provides the cross-video semantic alignment that automated retrieval alone cannot reliably ensure. For each video group, annotators author the question, answer, and evidence annotation. When the supporting evidence is localized, they mark the relevant temporal regions; when it depends on holistic narrative or causal understanding, they assign a global-semantic tag indicating that support is distributed across the video group rather than confined to a single interval. This scheme grounds localized Comparative Analysis and Tracking and Retrieval items while accommodating higher-order Integrated Reasoning questions. We then use LLMs to draft plausible distractors and polish question phrasing, with all edits verified by annotators; each final item has four answer options. LLM assistance is restricted to wording and distractor construction rather than cross-video QA generation, which is prone to hallucinated relations. As a final shortcut probe, each candidate is tested with compact open-source VLMs. Items solved by these limited-access models are returned to annotators for targeted revision, since they often contain shortcuts such as text-only answerability, option leakage, weak distractors, or memorized cues. Revised items re-enter the filter until they pass (Figure 2). Further details on the data construction and quality control processes are provided in Appendix B.

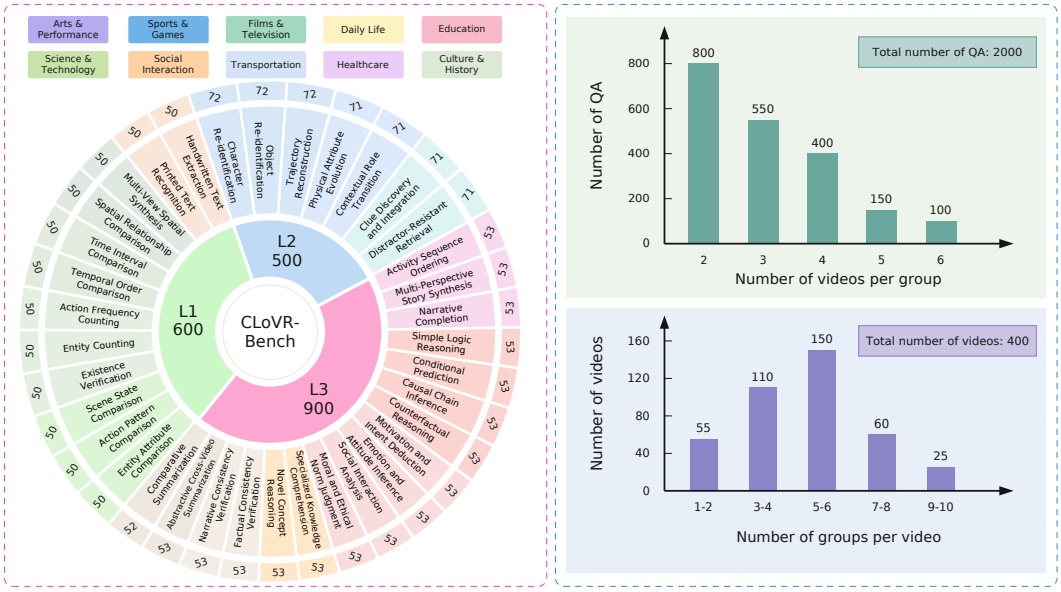

Figure 3: CLoVR-Bench statistics and task categories. It comprises 36 subtasks across diverse video domains; the outer ring gives the number of QA items per subtask. We also present the number of videos per group and the number of groups each video appears in.

4 HOLMES: Hypothesis-driven Observation, Linking, and Multi-source Evidence Synthesis

Cross-long-video reasoning has two separable operations: locating sparse decisive moments inside long videos, and linking the resulting evidence across streams. Collapsing these operations into option-conditioned search risks turning localization into answer confirmation, while collapsing videos into summaries risks inventing correspondences that were never visually checked. HOLMES treats finding and linking as separate stages (Figure 4; Algorithm 1). Given a question and candidate answers, the **Detective** plans evidence slots that discriminate among options without committing to any one of them: compact records of what visual evidence to find and what attribute to read after localization. The **Observer** fills these slots from the videos, writing fact cards when evidence is found and coverage certificates when a search fails. The **Cross-Examiner** links evidence across videos through side-by-side visual checks. The **Tribunal** reads the resulting graph and returns an answer only when support, binding validity, consistency, and de-duplication checks pass. The shared state is the *Case Dossier* $\mathcal{D}$, a typed graph that stores localized evidence, coverage records, cross-video bindings, constraints, and provenance across the loop.

4.1 Detective: Evidence-Slot Planning

The Detective transforms a question q and its candidate options into a hypothesis set $\mathcal{H}$, whose elements are the answer options, and a set of typed evidence slots $\{s_k\}$ whose collective filling will discriminate among them. The Detective sees no visual input, so no unverified visual claim can leak into the slot space. Each slot searches for an event whose occurrence can be verified without choosing an answer; readout attributes such as color, count, or identity are filled only after localization. The same localized event can update all options without option-conditioned search. Standard slots carry a state envelope: visual states expected immediately before and after the event. The event, pre-state, or post-state can trigger localization, allowing coarse anchors to bracket events they do not directly capture. Slots are typed as INTRA, LINKAGE, AGGREGATE, or NARRATIVE; linkage slots wait for dependency fact cards, and narrative slots are reserved for chapter-scale support. Risk mode follows a fixed rubric: FINEDETAIL for text or small/transient cues, NARRATIVE for causal or story-state evidence, and STANDARD otherwise.

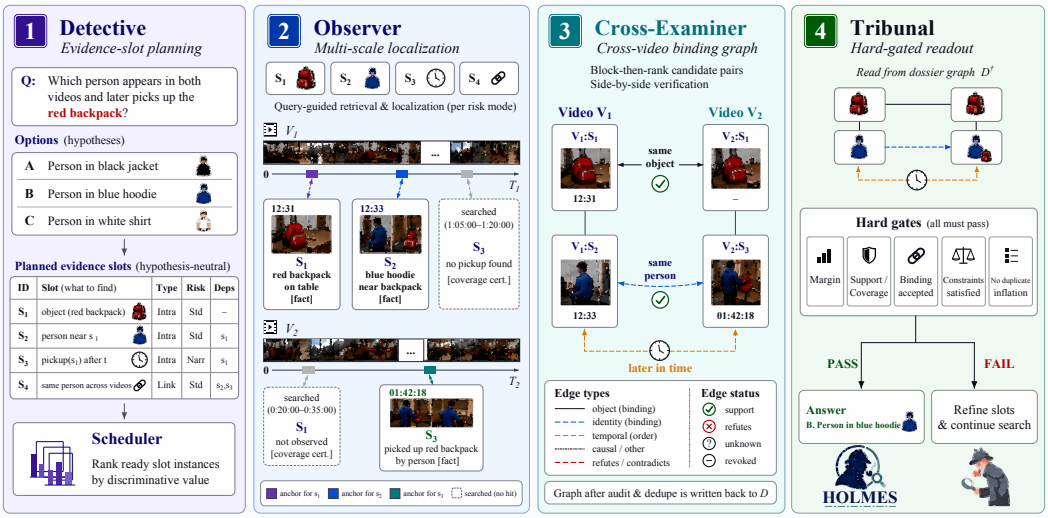

Figure 4: HOLMES Workflow. The four-stage agentic framework consists of the Detective, Observer, Cross-Examiner, and Tribunal, with a shared Case Dossier carrying evidence across stages.

Scheduling separates which slot to execute from how it should be localized. Priority is determined by the discriminative value of executing a slot instance:

$$\pi(\omega) \;=\; \frac{\mathrm{Disc}(s_k, \mathcal{H}, \mathcal{D}) \cdot \big(1 + \mathrm{Debt}(\omega)\big)}{\mathrm{Cost}(\omega)}, \tag{1}$$

where Disc, Debt, and Cost are deterministic functions of slot metadata and the current dossier (Appendix I). Intuitively, Disc favors slots that separate active hypotheses, Debt revisits under-covered evidence, and Cost penalizes expensive visual searches. The MLLM proposes semantic slots and discriminators; the controller only schedules from those records. Risk mode is then used separately to dispatch each selected instance to its localization strategy and coverage policy.

4.2 Observer: Multi-Scale State-Transition Localization

The Observer is the only module with frame-level access. For each scheduled slot instance, it dispatches to the localization strategy selected by the slot's risk mode and emits either fact cards or coverage certificates. It never emits an answer. In standard mode, the Observer first labels coarse anchors with local state relative to the slot predicate, then performs between-bracket refinement: compatible pre/post anchors open the transition interval, ambiguous anchors open local zoom windows, and long events are merged or split by localized readout. The key allocation rule is simple: spend frames where the state change must have occurred, not around isolated anchors. Fine-detail mode skips bracketing and uses denser scans plus crops for text, small objects, or transient cues; narrative mode inspects scene segments and emits semantic cards for distributed story state. Semantic cards are secondary evidence: localized fact cards, binding edges, and constraint edges take precedence whenever available. When localization produces no fact or semantic card, the Observer issues a coverage certificate rather than a negative fact. Coverage is computed from the search trace, not emitted as a subjective MLLM score (Appendix H). We enforce an absence invariant: only sufficiently covered certificates may support absence-dependent hypotheses; lower coverage creates scheduling debt for a denser recall-ladder pass. For linkage slots, a fact card in one video opens constrained counterpart search in the other referenced videos.

4.3 Cross-Examiner: Cross-Video Binding Graph

The Cross-Examiner makes cross-video binding an explicit operation on $\mathcal{D}$, rather than asking $\mathcal{M}$ to infer alignments from pooled textual context. Candidate binding pairs are first filtered by relation-specific hard feasibility constraints, such as temporal compatibility when applicable; entity signatures and scene context are used only as soft priors. The remaining pairs are ranked to prioritize comparisons whose acceptance or rejection could affect the active hypotheses, and the top pairs are

Algorithm 1 HOLMES inference loop

Require: Question q, candidate options $\mathcal{A}$, videos $\mathcal{V}$, MLLM $\mathcal{M}$, budgets $\mathcal{B}$

$(\mathcal{H}, \{s_k\}) \leftarrow \textsc{Detective}_{\mathcal{M}}(q, \mathcal{A})$ ▷ slot planning; no video access
$\mathcal{D} \leftarrow \emptyset$
for each investigation round within $\mathcal{B}$ **do**
  $\mathcal{Q} \leftarrow$ rank ready slot instances by Eq. 1, dispatch by risk mode, and truncate according to $\mathcal{B}$
  $(\mathcal{F}, \mathcal{C}) \leftarrow \textsc{Observer}_{\mathcal{M}}(\mathcal{Q}, \mathcal{V})$ ▷ fact/semantic cards $\mathcal{F}$, coverage certs $\mathcal{C}$
  $\mathcal{D}^{\dagger} \leftarrow \textsc{CrossExaminer}_{\mathcal{M}}(\mathcal{D} \cup \mathcal{F} \cup \mathcal{C}, \mathcal{H})$ ▷ block-then-rank, bind, audit, dedupe, tag
  $\{\sigma(h)\}_{h \in \mathcal{H}} \leftarrow$ Eq. 3 over $\mathcal{D}^{\dagger}$
  $\hat{h} \leftarrow \arg\max_{h \in \mathcal{H}} \sigma(h)$
  if $\textsc{GatesPass}(\hat{h}, \mathcal{D}^{\dagger})$ **then**
    return $\hat{h}$ with dossier provenance
  end if
  $g \leftarrow$ failed gates and bottlenecks from $\textsc{Tribunal}_{\mathcal{M}}$
  $\{s_k\} \leftarrow \textsc{Detective}_{\mathcal{M}}.\textsc{Refine}(\{s_k\}, \mathcal{D}^{\dagger}, g)$
  $\mathcal{D} \leftarrow \mathcal{D}^{\dagger}$
end for
return fallback over the latest dossier with incomplete-verification flag

shown to $\mathcal{M}$ side by side for the relevant predicate. Accepted comparisons become typed binding edges only after citing a discriminating visual cue, a consistency check, and a falsifier, and then surviving the constraint audit (Appendix I). Budget-pruned pairs remain UNKNOWN. Within $\mathcal{D}$, edges encode support, refutation, binding, temporal order, causal direction, and redundancy. The module audits hard constraints such as temporal compatibility, identity exclusivity, causal direction, and narrative contradictions. Edges whose falsifiers are triggered are revoked or downgraded, dependent stances are recomputed, and duplicate fact cards are collapsed before scoring (Appendix G).

4.4 Tribunal: Hard-Gated Graph Readout

The Tribunal closes each round by reading the answer out of the deduplicated graph $\mathcal{D}^{\dagger}$. It is a text-only MLLM call that drafts a readout record; the controller recomputes the score and gates below from $\mathcal{D}^{\dagger}$, and only this deterministic check decides whether an answer is returned. The readout separates investigation control from final commitment: a soft support-difference score identifies which hypotheses and slots remain underdetermined, while hard gates decide whether the graph can support an answer. The score is not a calibrated probability or a verdict rule; its formal expression is given in Appendix I. A hypothesis $\hat{h} = \arg\max_k \sigma(h_k)$ is returned as the verdict only when hard gates pass: sufficient margin, admissible positive support or coverage, accepted binding edges, no active constraint violation, and no duplicate inflation (formal gates in Appendix I). A hypothesis cannot win merely because alternatives are weakly refuted or because duplicated evidence is numerous. If any gate fails, the Tribunal returns the failing gate, underdetermined slots, missing discriminators, and bottleneck type (coverage, ambiguity, linkage, temporal consistency, or aggregation). The Detective refines only the implicated slots. If the investigation budget is exhausted, the Tribunal returns an evidence-grounded fallback with an incomplete-verification flag; for aggregate queries with incomplete coverage, it returns a count interval. The score guides investigation, while the hard gates determine whether the graph is strong enough to answer.

5 Experiments

We evaluate 13 MLLMs, covering closed-source GPT [21, 22] and Gemini [6] systems and open-source Qwen2.5-VL [3], Qwen3-VL [2], InternVL2.5 [5], InternVL3 [42], and LLaVA-OneVision [12] families. Following MVU-Eval [24], each video receives a uniform 32-frame budget. Frames are grouped by source video, marked with explicit video identifiers, and placed before the question and answer options in a single multimodal context. Unless a model requires a different interface, frames are resized with aspect ratio preserved and the longer side capped at 720 pixels. For models with specific resolution requirements, we follow the recommended input resolution while keeping the same 32-frame-per-video budget. The 720-pixel cap applies to this protocol; denser settings trade

Table 2: Performance on CLoVR-Bench. Task abbreviations are defined in Appendix A. Blue marks the best open-source model, and Red marks the best closed-source model. Numbers in (red) report absolute improvement over base model. The best model result in each column is in **bold**.

Models	Overall	Comparative Analysis					Tracking & Retrieval			Integrated Reasoning					
		ASC	EC	TA	SA	OCR	CIA	SET	PDR	ER	CR	PSR	DKA	MSCV	IS
Human	88.6	92.4	85.1	80.3	91.6	94.2	90.5	87.8	84.6	89.4	93.1	86.2	92.7	83.5	88.8
Closed-Source Models															
GPT-4o [21]	38.3	48.3	40.5	34.2	52.6	56.8	41.5	38.8	32.6	31.7	37.2	24.5	36.4	26.3	34.8
GPT-4.1 [22]	41.0	51.8	43.7	37.5	55.9	60.4	44.2	41.6	35.5	34.3	39.8	26.4	38.5	28.2	36.6
Gemini-2.0-Flash [6]	36.0	45.6	37.8	31.5	49.7	53.4	39.3	36.7	30.5	30.1	35.4	22.8	34.2	24.5	32.6
Gemini-2.5-Pro [6]	44.3	54.2	47.4	41.1	59.8	64.5	47.3	44.6	38.2	37.5	43.2	29.8	41.3	31.5	39.7
Open-Source Models <10B															
Qwen2.5-VL-7B [3]	26.8	33.6	26.8	22.4	37.1	39.5	29.3	26.4	22.5	23.5	27.6	16.7	26.4	18.7	24.6
Qwen3-VL-8B [2]	32.0	40.5	32.7	27.4	44.2	46.8	34.5	31.6	26.5	27.5	32.4	20.6	31.5	22.7	29.4
InternVL2.5-8B [5]	29.1	36.5	29.4	24.6	40.2	42.8	31.5	28.6	24.3	25.4	29.7	18.4	28.6	20.5	26.8
InternVL3-8B [42]	31.0	39.3	31.5	26.4	42.7	45.5	33.4	30.5	25.7	26.7	31.5	19.8	30.6	21.9	28.5
LLaVA-OneVision-7B [12]	26.4	33.2	26.5	22.3	36.4	38.8	28.7	26.1	22.4	23.4	27.2	16.4	25.8	18.3	24.2
Open-Source Models >70B															
Qwen2.5-VL-72B [3]	36.6	45.8	38.4	32.1	49.6	52.8	39.3	36.7	30.8	31.6	36.5	23.9	35.4	25.7	33.6
InternVL2.5-78B [5]	34.3	42.6	35.7	29.8	46.4	49.5	36.5	34.2	28.7	29.5	34.6	22.4	33.7	24.3	31.8
InternVL3-78B [42]	37.9	47.2	39.6	33.8	51.5	54.7	40.7	37.9	32.4	32.4	37.6	24.7	36.5	26.8	34.5
LLaVA-OneVision-72B [12]	33.0	41.2	33.8	28.4	44.7	47.6	35.4	32.7	27.5	28.4	33.5	21.6	32.5	23.4	30.6
HOLMES (Qwen2.5-VL-7B)	41.0 (+14.2)	42.5 (+8.9)	40.2 (+13.4)	37.5 (+15.1)	46.8 (+9.7)	49.4 (+9.9)	43.6 (+14.3)	42.5 (+16.1)	37.4 (+14.9)	39.5 (+16.0)	43.7 (+16.1)	30.4 (+13.7)	42.5 (+16.1)	37.8 (+19.1)	40.6 (+16.0)
HOLMES (Gemini-2.5-Pro)	**54.8** (+10.5)	**59.4** (+5.2)	**55.6** (+8.2)	**52.7** (+11.6)	**63.5** (+3.7)	**67.4** (+2.9)	**61.5** (+14.2)	**58.7** (+14.1)	**50.6** (+12.4)	**50.4** (+12.9)	**54.8** (+11.6)	**41.6** (+11.8)	**53.4** (+12.1)	**46.5** (+15.0)	**51.7** (+12.0)

spatial resolution for temporal coverage (Appendix K.2). Additional model evaluations are reported in Appendix K.1.

5.1 CLoVR-Bench Evaluation Results and Analysis

Table 2 reports per-task accuracies for 13 models and two HOLMES instantiations, one on a small open backbone and one on a strong closed backbone. Four patterns stand out. First, accuracy drops monotonically from L1 to L3 for every base model, showing that long-range evidence search (L2) and cross-video synthesis (L3) are genuinely harder than localized perception (L1), not artifacts of weaker recognition. Second, three subtasks remain hard regardless of model scale: TA in L1, where temporal reasoning is hard even when intervals are given; and PSR and MSCV in L3, which require narrative grounding and explicit cross-video binding that current MLLMs do not handle well. CR and DKA are the easiest L3 subtasks because they can be solved by the LLM's commonsense once a few visual cues are extracted, so the L3 average should be read together with MSCV, PSR, and IS when comparing methods. Third, model scale helps L1 more than L3, and the weakest 7B models fall below 25 on PSR and MSCV, suggesting systematic option biases and evidence misses rather than uniformly random guessing. Humans stay above 80 on every subtask. For every model, the largest model–human gap is on PSR, and the next largest gaps are also L3 subtasks (ER, MSCV, DKA, or CR); TA shows a smaller gap because it is also the hardest subtask for humans. Fourth, HOLMES improves both backbones, with the smallest gain on L1 and the largest gain on MSCV. This pattern is consistent with the design: L1 questions already provide timestamps so finding adds little, while L2 and cross-video L3 questions require the explicit finding and linking that HOLMES introduces.

Difficulty-factor analysis. To identify which data factors most affect cross-long-video reasoning, we measure Qwen2.5-VL-7B accuracy under four conditions. Duration and Evidence sparsity partition all QA pairs into disjoint subsets, # Videos contrasts two-video queries with queries over four or more videos, and Frames per video reruns the full benchmark at three sampling budgets. Evidence sparsity is determined from the annotated supporting intervals: Dense (>10% coverage and visually salient), Sparse (≥1% and ≤10% coverage with moderate salience), and Needle-like (<1% coverage or visually subtle). Three patterns emerge from Table 3. First, adding frames improves accuracy, but with rapidly diminishing re-

Table 3: Sensitivity analysis on CLoVR-Bench.

Factor	Setting	Acc.	Δ
Frames per video	32	26.8	–
	64	28.5	↑ +1.7
	128	29.2	↑ +2.4
# Videos	2	30.4	–
	4+	22.6	↓ -7.8
Duration per video	30 to 60 min	29.5	–
	60+ min	23.7	↓ -5.8
Evidence sparsity	Dense	33.6	–
	Sparse	26.5	↓ -7.1
	Needle-like	19.8	↓ -13.8

Table 4: Comparison with representative long-video baselines on CLoVR-Bench. All methods use Qwen2.5-VL-7B and a 32-frame/video external source-frame cap; cumulative tokens and calls are not matched. The best result in each column is in **bold**.

Method	Comparative Analysis	Tracking & Retrieval	Integrated Reasoning	Overall
Keyframe-Selection Based				
AKS [32]	35.4	29.7	26.5	30.5
FOCUS [44]	36.8	31.2	27.6	31.8
Q-Frame [40]	38.1	32.6	28.9	33.0
Agent-Based				
VideoAgent [7]	39.5	34.2	30.4	34.5
LVAgent [4]	40.4	36.6	32.9	36.4
HOLMES (Ours)	**43.3**	**41.2**	**39.1**	**41.0**

turns. Second, accuracy drops sharply on 4+ video and 60+ min queries under the same per-video frame budget, indicating that the bottleneck is cross-video and long-context interference rather than per-video coverage. Third, evidence sparsity has the largest effect, with Needle-like cases falling below random accuracy. This suggests that when the model cannot locate the key evidence, it tends to commit to plausible distractors rather than guess uniformly. More frames mitigate visual undercoverage but do not resolve the finding and linking failures in sparse and multi-video queries, which is the gap HOLMES targets.

5.2 Evaluating the Effectiveness of HOLMES

No prior method specifically targets cross-long-video reasoning, so we adapt two families of representative long-video methods to this setting and compare them with HOLMES under the same Qwen2.5-VL-7B backbone and a ceiling of 32 unique source frames per video. This ceiling does not match cumulative visual tokens or model calls. Resource accounting and comparisons with common token and call ceilings are given in Appendices K.3 and K.4. Table 4 shows that performance grows in the order keyframe selection $<$ video agents $<$ HOLMES, with HOLMES achieving the best result on every level. The lead is smallest on Comparative, where existing methods already handle per-video evidence finding when intervals are given, and grows on Tracking and Integrated, where locating sparse evidence under cross-video interference and binding evidence across videos become the bottleneck.

Generalization to long-video benchmarks.

CLoVR-Bench targets cross-long-video reasoning, but HOLMES's finding-and-linking design is generic and should also help on single-video long-form benchmarks where evidence localization alone is the main bottleneck. We therefore apply HOLMES as an inference-time wrapper on top of two open-source base models and evaluate on LongVideoBench (validation accuracy) [37] and the MLVU dev set (M-Avg over its 7 multiple-choice tasks) [41]. HOLMES improves both backbones on both benchmarks (Table 5). The gains are smaller than on CLoVR-Bench, as expected: neither benchmark stresses cross-video binding, so only the finding stage of HOLMES is fully exercised. The improvement is nonetheless consistent across two distinct backbones and two benchmark suites, indicating transfer beyond the proposed taxonomy. Under common frame, token, and call ceilings, the HOLMES–LVAgent confidence intervals on these two single-video benchmarks include zero; the multi-video MVU-Eval gain remains significant (Appendix K.5).

Table 5: Generalization to video benchmarks.

Method	LongVideoBench	MLVU
Qwen2.5-VL-7B	54.7	70.2
+ **HOLMES**	**58.4** (+3.7)	**72.5** (+2.3)
InternVL3-8B	62.3	71.4
+ **HOLMES**	**65.5** (+3.2)	**73.6** (+2.2)

5.3 Ablation Studies and Error Analysis

We replace each stage of the finding-linking-readout loop with a simpler alternative, keeping the base model and budget fixed. Replacing visual binding with summary-only linking causes the largest drop, since cross-video binding is the operation no other stage can recover. Uniform frame search and answer-conditioned planning come next, both targeting the finding stage. Removing the hard

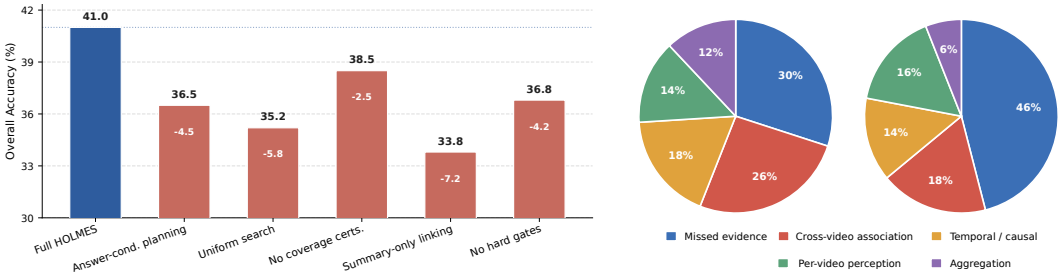

Figure 5: Ablation and error analysis on CLoVR-Bench. Left: overall accuracy of HOLMES variants. Right: error distributions for HOLMES (left pie) and the uniform-sampling baseline (right pie).

readout gates produces a moderate drop concentrated on high-confusion options, while removing coverage certificates is the mildest, consistent with this stage governing only absence-dependent items. The four stages are therefore complementary rather than redundant, and linking is the single most consequential stage for cross-long-video reasoning.

We manually inspect 200 incorrect predictions for HOLMES and for the uniform-sampling baseline, both built on Qwen2.5-VL-7B (Figure 5). The baseline is dominated by missed evidence, which alone accounts for 46% of its errors. HOLMES cuts this share to 30% through risk-conditioned localization, while the share of cross-video association errors rises from 18% to 26%, close behind missed evidence. This shift is consistent with the ablation: as finding errors decline, linking accounts for a growing share of the remaining errors. Per-video perception and temporal/causal reasoning together account for roughly one-third of HOLMES errors and indicate that future gains lie in tighter visual verification within localized regions, not only in expanding retrieval coverage.

6 Conclusion

We introduced CLoVR-Bench, a comprehensive benchmark for reasoning across multiple long-form videos, organized by a three-level taxonomy over 14 tasks and 36 subtasks. Our evaluation shows that current MLLMs struggle not only because videos are long, but also because decisive evidence is sparse and must be explicitly linked across streams. We further proposed HOLMES, a training-free framework that turns cross-long-video QA from monolithic prompting into structured evidence acquisition, visual binding, and hard-gated graph readout over an auditable Case Dossier. Our results suggest that future progress on cross-long-video reasoning will require systems that can both search selectively and verify cross-video relations, rather than merely scale frame budgets or summarize longer contexts. Together, CLoVR-Bench and HOLMES provide a diagnostic testbed and an interpretable inference baseline for this direction.

Limitations. The benchmark covers multiple-choice QA over public videos. HOLMES adds inference cost that grows with the number and length of input videos (Appendix K.6), and it only partially recovers from retrieval and binding errors (Appendix K.7). Human solvers view complete videos without a frame cap, so model–human gaps are not measured under matched viewing budgets. We discuss these limitations and the release terms further in Appendix E.

Acknowledgments and Disclosure of Funding

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). The authors declare no competing interests.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, et al. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. LVAgent: Long video understanding by multi-round dynamical collaboration of MLLM agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20237–20246, 2025.

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

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[7] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024.

[8] 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*, pages 24108–24118, 2025.

[9] Google DeepMind. Gemini 3.1 Pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

[10] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-LLM based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712, 2025.

[11] De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. FRAG: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*, 2025.

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[13] Jingyao Li, Jingyun Wang, Molin Tan, Haochen Wang, Cilin Yan, Likun Shi, Jiayin Cai, Xiaolong Jiang, and Yao Hu. CrossVid: A comprehensive benchmark for evaluating cross-video reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6244–6252, 2026.

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

[15] Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. KeyVideoLLM: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*, 2024.

[16] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. PanoSent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7667–7676, 2024.

[17] Meng Luo, Bobo Li, Shanqing Xu, Shize Zhang, Qiuchan Chen, Menglu Han, Wenhao Chen, Yanxiang Huang, Hao Fei, Mong-Li Lee, et al. Unveiling the cognitive compass: Theory-of-mind-guided multimodal emotion reasoning. *arXiv preprint arXiv:2602.00971*, 2026.

[18] Meng Luo, Yanlin Li, Hao Li, Hongzhan Lin, Pengfei Zhou, Tianjie Ju, Ran Zhang, Yeying Jin, Mong-Li Lee, and Wynne Hsu. AI for games in the foundation model era. *arXiv preprint arXiv:2609.16679*, 2026.

[19] Meng Luo, Yicheng Liu, Jiahao Wang, Yuanxing Zhang, Xin Tao, Pengfei Wan, Kun Gai, and Hao Fei. From evaluation to enhancement: Benchmarking and improving think-with-video reasoning for video generative models. In *European Conference on Computer Vision*, pages 546–563. Springer, 2026.

[20] Meng Luo, Shengqiong Wu, Liqiang Jing, Tianjie Ju, Li Zheng, Jinxiang Lai, Tianlong Wu, Xinya Du, Jian Li, Siyuan Yan, et al. Dr. V: A hierarchical perception-temporal-cognition framework to diagnose video hallucination by fine-grained spatial-temporal grounding. *International Journal of Computer Vision*, 134(6):278, 2026.

[21] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[22] OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, 2025.

[23] OpenAI. GPT-5.1: A smarter, more conversational ChatGPT. `https://openai.com/index/gpt-5-1/`, 2025.

[24] Tianhao Peng, Haochen Wang, Yuanxing Zhang, Zekun Wang, Zili Wang, Gavin Chang, Jian Yang, Shihao Li, Yanghai Wang, Xintao Wang, et al. MVU-Eval: Towards multi-video understanding evaluation for multimodal LLMs. *arXiv preprint arXiv:2511.07250*, 2025.

[25] Qwen Team. Qwen3.5-9B model card. `https://huggingface.co/Qwen/Qwen3.5-9B`, 2026.

[26] Qwen Team. Qwen3.6-35B-A3B model card. `https://huggingface.co/Qwen/Qwen3.6-35B-A3B`, 2026.

[27] Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. *arXiv preprint arXiv:2601.18157*, 2026.

[28] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-augmented generation with extreme long-context videos. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2390–2401, 2026.

[29] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169, 2025.

[30] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[31] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A training-free approach for list-wise frame selection in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24090–24101, 2025.

[32] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128, 2025.

[33] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[34] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[35] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. LongVLM: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024.

[36] Chao-Yuan Wu and Philipp Krähenbühl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021.

[37] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[38] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

[39] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in MLLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12000–12008, 2026.

[40] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-Frame: Query-aware frame selection and multi-resolution adaptation for Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22056–22065, 2025.

[41] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701, 2025.

[42] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[43] Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, Xiongfei Yao, and Shuaiwei Jiao. CVBench: Benchmarking cross-video synergies for complex multimodal reasoning. *arXiv preprint arXiv:2508.19542*, 2026.

[44] Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. *arXiv preprint arXiv:2510.27280*, 2025.

[45] Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024.

Appendix

In the appendix, we provide definitions of all CLoVR-Bench tasks and subtasks (Section A), the complete dataset construction and quality-control procedure (Section B), human baseline and annotation-reliability details (Section C), the HOLMES implementation protocol and responsible-release discussion (Section D– E), the module prompts and output schemas (Section F), the Case Dossier schema (Section G), coverage functions (Section H), scheduling, localization, and binding details (Section I), and the default HOLMES budget policy (Section J). Additional evaluations, resource accounting, budget-controlled comparisons, and robustness analyses are provided in Section K.

A Task Definitions

This appendix provides the formal definitions of all 36 subtasks instantiated under our three-level taxonomy. Each definition specifies the cognitive focus of the subtask and the principal capability it is designed to probe.

A.1 Comparative Analysis

A.1.1 Attribute and State Comparison (ASC)

Entity Attribute Comparison. Compares static visual attributes of entities across videos, including color, shape, size, and other directly observable physical properties. Given two or more localized clips, the model must qualitatively contrast these attributes between corresponding entities, with no requirement for inference about temporal evolution or causal relationships.

Action Pattern Comparison. Evaluates the comparison of dynamic single-agent action patterns across videos, focusing on execution style, technique, and kinematic characteristics (e.g., the entry posture of two divers). To disambiguate from multi-entity interaction analysis, the comparison is restricted to behaviors performed by a single individual or to symmetric behaviors performed independently by separate agents.

Scene State Comparison. Targets holistic environmental contrast, including ambient lighting, tidiness, weather conditions, and overall atmosphere. Unlike fine-grained attribute comparison, this subtask emphasizes the gestalt impression of a scene, requiring models to integrate diffuse visual cues into a coherent qualitative judgment.

A.1.2 Existence and Counting (EC)

Existence Verification. Probes whether models can confirm or deny the presence of a specified target (an entity, attribute, or action) within designated portions of given videos. The verification is binary in nature and requires careful visual inspection rather than inference.

Entity Counting. Evaluates static counting of distinct entities (e.g., people, vehicles, animals) within and across video segments. Models must perform accurate enumeration and either provide totals or compare counts between videos.

Action Frequency Counting. Targets dynamic counting of repetitive actions or events (e.g., the number of squats performed). The frequency comparison may be intra-video or cross-video, but never requires causal interpretation of why such events occur.

A.1.3 Temporal Analysis (TA)

Temporal Order Comparison. Compares the relative ordering of events between two or more videos using only surface-level timing information. Crucially, the events compared are designed to be causally independent (e.g., distinct news segments in a broadcast), explicitly avoiding cases where temporal order implies causation, which would belong to Level 3.

Time Interval Comparison. Evaluates the contrast of durations or inter-event intervals across videos, such as which of two athletic routines lasted longer. Judgments require only direct timing observation, not reasoning about why one interval differs from another.

A.1.4 Spatial Analysis (SA)

Spatial Relationship Comparison. Focuses on comparing directly perceivable spatial relationships across videos, such as the relative positions, orientations, and layouts of objects or agents. Typical instances involve before–after layout changes (e.g., furniture rearrangement) viewed from a consistent perspective.

Multi-View Spatial Synthesis. Requires the integration of complementary spatial information from videos recorded at different viewpoints to recover details that are occluded in any single view (e.g., inferring an object hidden in a frontal shot via a side-rear shot). It tests the ability to fuse multi-perspective evidence into a unified spatial representation.

A.1.5 Optical Character Recognition (OCR)

Printed Text Recognition. Evaluates the recognition and semantic interpretation of standardized printed text appearing in video scenes (e.g., signage, posters, slide content). Although the visual modality differs from coarse object recognition, the cognitive demand remains at the perception level.

Handwritten Text Extraction. Evaluates the recognition of non-standard handwritten content (e.g., blackboard derivations, sticky notes, manuscripts), introducing additional challenges of legibility and stylistic variance compared to printed text.

A.2 Tracking and Retrieval

A.2.1 Cross-Video Identity Association (CIA)

Character Re-identification. Requires models to identify the same person across different videos despite variations in clothing, environment, lighting, or viewpoint. The model must maintain a stable identity representation that persists through such surface-level changes.

Object Re-identification. Analogous to character re-identification but applied to inanimate objects, this subtask tracks an object's reappearance and circulation across long videos. Negative cases (i.e., the object does not reappear) are deliberately included to prevent shallow heuristic matching.

Trajectory Reconstruction. Requires the integration of multiple long videos—typically from disjoint camera feeds—to reconstruct the complete movement path of a person or vehicle. It tests the ability to chain together fragmented spatiotemporal evidence into a coherent trajectory.

A.2.2 State Evolution Tracking (SET)

Physical Attribute Evolution. Tracks gradual or abrupt changes in observable physical attributes of an entity (e.g., a character's hairstyle and facial appearance, a glacier's coverage area) across temporally separated videos. The model must identify the trajectory of change rather than merely the endpoints.

Contextual Role Transition. Follows changes in an entity's functional role across different contexts (e.g., a bottle of whisky used as a beverage in one scene and as an antiseptic in another). Such transitions are inferred from observed usage patterns rather than from explicit annotations, requiring contextual sensitivity.

A.2.3 Precision Detail Retrieval (PDR)

Clue Discovery and Integration. Requires the model to recover sparse, easily overlooked details from different long videos and combine them to produce the answer (e.g., a username uttered in one video and an email domain shown in another). It places high demands on long-context memory and cross-video information binding.

Distractor-Resistant Retrieval. Evaluates retrieval robustness when the queried target is accompanied by visually similar distractors (e.g., two red sports cars with different license plates appearing in different videos). The model must precisely localize the queried instance without confusing it with the distractor.

A.3 Integrated Reasoning

A.3.1 Event Reconstruction (ER)

Activity Sequence Ordering. Requires the model to recover the original procedural ordering of activities depicted across multiple videos (e.g., washing $\rightarrow$ cutting $\rightarrow$ cooking vegetables). It tests procedural reasoning over related but temporally disjoint events.

Multi-Perspective Story Synthesis. Synthesizes multiple long videos that document the same complex event from different viewpoints (e.g., body-cam footage and bystander recordings) into a single coherent narrative. The model must reconcile partial and potentially complementary observations to select the account best supported by the available videos.

Narrative Completion. Requires the model to infer missing intermediate events given only the preceding and following narrative segments (e.g., the protagonist trapped in a cave at the end of one video appears safe at home in the next). This bridging inference draws on both observed evidence and plausibility reasoning.

A.3.2 Causal Reasoning (CR)

Simple Logic Reasoning. Probes single-step cause–effect inferences spanning videos (e.g., a stain in one video is directly attributable to a spill in another). It tests cross-video causal grounding without requiring a multi-hop causal chain.

Conditional Prediction. Presents an explicit rule, warning, or physical condition in one video and a triggering action in another, requiring the model to predict the necessary consequence according to the established condition. It tests rule application across video boundaries.

Causal Chain Inference. Demands recovery of multi-hop causal chains in which an outcome arises from a sequence of intermediate causes (e.g., embezzlement $\rightarrow$ substandard construction $\rightarrow$ cursory inspection $\rightarrow$ extreme rainfall $\rightarrow$ bridge collapse). The model must trace the chain back to its root cause rather than stopping at proximate ones; distractor options correspond to non-root links to penalize shallow attribution.

Counterfactual Reasoning. Perturbs an established condition in one video and asks the model to deduce how outcomes in subsequent videos would change (e.g., if the protagonist had not encountered the doctor in a prior video, which downstream events would no longer occur). It tests reasoning over hypothetical alternative worlds while preserving causal consistency.

A.3.3 Psychological and Social Reasoning (PSR)

Motivation and Intent Deduction. Infers the underlying intentions behind characters' actions by integrating behavioral evidence with prior contextual setup (e.g., a tactical plan revealed in an earlier video explains an apparently irrational order on the battlefield).

Emotion and Attitude Inference. Draws on both observed expressions and historical relational context to infer complex affective states. Surface-level cues are often insufficient or misleading without context from additional videos (e.g., a stoic coach's tear interpreted as pride rather than sorrow).

Social Interaction Analysis. Examines group dynamics, role assignments, and collaborative behaviors within social events (e.g., distinguishing the planner from the executor within a crew during a coordinated operation). It requires inference over multi-agent interaction patterns.

Moral and Ethical Norm Judgment. Evaluates whether observed behaviors violate stated or widely recognized ethical, professional, or social norms (e.g., a physician breaching patient confidentiality by sharing case details socially). The item is grounded in the depicted actions and context rather than in an open-ended value debate.

A.3.4 Domain Knowledge Application (DKA)

Specialized Knowledge Comprehension. Requires the application of established domain expertise (e.g., physical laws, biological mechanisms) to interpret phenomena observed in the videos. The model must map abstract principles to concrete visual events (e.g., explaining a racing accident

via Bernoulli's principle introduced in a companion video), rather than only checking whether two sources are mutually consistent.

Novel Concept Reasoning. Emphasizes on-the-fly learning of newly introduced—and often counter-intuitive—concepts presented in one video, then applying them to interpret another (e.g., a fictional culture in which nodding signifies refusal). It evaluates rapid knowledge acquisition and disciplined suppression of contradictory pretraining priors, rather than rote recall of pre-existing knowledge.

A.3.5 Multi-Source Consistency Verification (MSCV)

Factual Consistency Verification. Checks whether claims or depictions across the videos are consistent with established real-world facts, requiring the model to detect anachronisms or factual contradictions (e.g., an AK-47 appearing in a 1942 Stalingrad setting). Unlike Domain Knowledge Application, the task is verification of consistency, not explanation by applying a specialized principle.

Narrative Consistency Verification. Checks internal narrative coherence within fictional content, where external truth is irrelevant. The model must detect self-contradictions among claims and depictions across videos (e.g., a character claims never to have visited Paris, yet appears beneath the Eiffel Tower in another video).

A.3.6 Integrated Summarization (IS)

Abstractive Cross-Video Summarization. Requires judging which candidate summary best fuses complementary information from multiple long videos with different focal points (e.g., a launch keynote, an independent review, and a market analysis). The selected summary must cover the central evidence without introducing unsupported claims.

Comparative Summarization. Emphasizes contrast rather than fusion. The model must identify which candidate summary best captures divergences in narrative stance, focus, or framing across videos covering the same event (e.g., an official documentary versus an independent investigative piece). It tests recognition of perspective and framing while remaining within the multiple-choice format.

B Dataset Construction

This appendix expands on the construction pipeline, providing procedural detail, model configurations, and quality-control criteria.

B.1 Video Sourcing

All source videos are collected from publicly available web platforms, predominantly YouTube, with supplementary material from other open video repositories accessed via Google search. We enforce a strict minimum duration of 30 minutes per video for the long-form evaluation setting. Rather than relying on automated retrieval alone (e.g., caption similarity or keyword matching), expert annotators select source groups. For each candidate QA, annotators identify videos whose joint content exhibits the spatiotemporal patterns, entity correspondences, or causal structures required by the corresponding subtask definition. We provide annotators with: (i) the target subtask's formal definition and exemplar QA pairs, (ii) a checklist of inter-video relations the group must support (e.g., shared entities for re-identification, temporal precedence for causal-chain inference, complementary perspectives for multi-view spatial synthesis), and (iii) a duration guideline ensuring no video falls below the 30-minute floor. This human-led sourcing step reduces topically related but relationally underspecified groups, which otherwise lead to trivial or unanswerable questions.

B.2 QA Authoring and Evidence Annotation

Once a video group is curated, annotators directly compose the question, the correct answer, and a set of supporting evidence annotations. We deliberately do not use LLMs for QA generation itself, because pilot inspection of LLM-generated cross-video QA pairs revealed two recurring failure modes: hallucinated cross-video relationships not present in the videos, and over-reliance on surface-

level patterns that single-video reasoning can already exploit. For evidence annotation, every QA pair is tagged with one of two annotation types:

- **Bounded evidence regions.** For QA pairs whose answers can be supported by specific video segments, annotators mark the timestamps in each relevant video where supporting evidence appears. A single QA may have multiple disjoint evidence intervals spanning multiple videos. These items predominantly populate Level 1 (Comparative Analysis) and Level 2 (Tracking and Retrieval) tasks.
- **Global-semantic items.** For QA pairs whose answers require holistic understanding, such as narrative arc, causal trajectory, thematic coherence, or social-psychological inference, and cannot be reduced to bounded intervals, annotators mark the item as *global-semantic*. These items predominantly populate Level 3 (Integrated Reasoning) tasks.

This dual annotation scheme is structurally aligned with our three-level taxonomy and enables auxiliary analyses such as temporal-grounding evaluation, evidence-region attention diagnostics, and per-level interpretability.

B.3 Distractor Generation and Phrasing Refinement

For multiple-choice items, plausible and confusable distractors are essential to benchmark difficulty. We use GPT-5.1 [23] and Gemini-2.5-Pro [6] as writing assistants for this stage, providing them with the question, correct answer, video-group metadata, and task definition. This stage is text-only: the writing assistants do not process videos, and GPT-5.1 is not an evaluation backbone. The models are prompted to: (i) generate three to five plausible distractors that share surface features with the correct answer, (ii) ensure distractors are factually incorrect when verified against the video content, and (iii) suggest minor phrasing refinements to remove ambiguity.

Annotators then review every model-suggested distractor and phrasing change. The reviewer confirms that the rewritten options preserve the original task intent, remain consistent with the formal subtask definition, and introduce genuine confusability rather than superficial variation. Suggestions that violate these criteria are rejected; suggestions that partially satisfy them are edited by the annotator before adoption. Each final item retains three distractors, giving four answer options. LLM involvement is limited to lexical and option-level refinement, where outputs remain easy to verify against ground truth, rather than to QA generation.

B.4 Small-VLM Leakage Filtering

For CLoVR-Bench, candidate items are checked for shortcuts from text, options, surface frames, and pre-training memorization. Items should require reasoning over multiple long-form videos rather than a single language-only or surface-level cue.

We address this with a small-VLM probe. Each candidate QA pair is presented to two compact open-source vision-language models, Qwen3-VL-2B [2] and InternVL3.5-2B [34], under the same input setting used at downstream evaluation: 32 frames per video sampled uniformly from the full duration, with frames resized so that the longer side is at most 720 pixels. This probe is used as a shortcut detector rather than as a measure of model capacity. Under sparse uniform sampling from several 30+ minute videos, compact VLMs have limited access to the decisive evidence and limited capacity for long multi-video binding. When either small VLM solves a candidate item, the item is flagged for revision. We classify the typical shortcut modes as:

- **Text-only answerability.** The question and options imply the answer without requiring video content.
- **Option-level leakage.** Options leak the correct answer through length asymmetry, lexical specificity, formatting cues, or distractor implausibility.
- **Distractor weakness.** Distractors are implausible enough that elimination yields the correct answer regardless of video content.
- **Pre-training memorization.** For items grounded in well-known films, public events, or widely discussed videos, the model may have memorized the answer from web text during pre-training.

Items flagged by either small VLM enter the revision loop described next.

B.5 Human Revision Loop and Final Quality Control

Flagged items are not discarded or auto-regenerated by an LLM. Instead, they are returned to the original annotator with the small-VLM diagnostic indicating the suspected shortcut category. The annotator performs targeted revision, such as rewriting the question, rebalancing distractors, or, in rare cases, replacing one or more source videos in the group, and resubmits the item. The item then re-enters the leakage filter. This loop iterates until the item passes both small-VLM probes.

After leakage filtering, every candidate undergoes a final independent human quality-control review by an annotator who did not author the original item. The reviewer verifies: (i) factual correctness of the marked answer against the videos, (ii) correctness and completeness of evidence-region annotations, (iii) genuine reliance on cross-video integration, meaning the item must be unanswerable from any single video alone, (iv) consistency with the formal subtask definition, and (v) absence of residual ambiguity. Items failing this review are returned for further revision; items passing it enter the final benchmark. The completed CLoVR-Bench contains 2,000 verified QA pairs grounded in 400 long-form videos, each reviewed through human quality control and a multi-model small-VLM shortcut probe.

B.6 Group and Duration Statistics

Items are balanced across subtasks: each of the 12 Comparative Analysis subtasks has 50 items, the 7 Tracking and Retrieval subtasks have 71–72 items each, and the 17 Integrated Reasoning subtasks have 52–53 items each, giving 600, 500, and 900 items per level (Figure 3). The 400 videos form 512 source groups, with a median of four QA pairs per group (interquartile range: 3–5). The mean number of videos per group is 3.5 when groups are weighted equally and 3.1 when weighted by QA count. Queries contain two to six videos: 800 queries use two videos, 950 use three or four, and 250 use five or six. Video duration has minimum, median, 90th percentile, and maximum values of 30.0, 47.8, 76.4, and 118.7 minutes, respectively. These statistics distinguish source-group counts from the number of QA queries and source-video instances.

C Human Baseline and Annotation Reliability

Annotator pool and expertise. CLoVR-Bench was produced by eight expert annotators with prior experience in video QA annotation, multimodal dataset auditing, or domain-specific video analysis. Before contributing final items, annotators completed a calibration round covering all three taxonomy levels, evidence-interval marking, cross-video-link annotation, and ambiguity rejection. Calibration examples were not included in the benchmark. Annotators worked from written task definitions and checklists, and compensation followed the higher of the applicable institutional rate or local minimum-wage requirement.

Human row in Table 2. The reported human accuracy is not computed from the authors of the QA items and does not use gold evidence intervals. Solvers were drawn from the annotator pool with item-level conflict blocking. Each final item was assigned to a solver who had not authored or quality-controlled that item. Solvers received the same question and options as models and could inspect the complete source videos with pause, rewind, and manual scrubbing, but they did not see the annotated supporting intervals, cross-video links, or the correct answer. We use this protocol for the human row in Table 2.

Overlap and agreement. To estimate reliability, 600 items stratified by level, task family, number of videos, and evidence sparsity received two additional blind solves, yielding three independent answers per item. Raw three-way answer agreement on this audit split was 86.9%, and Fleiss' κ was 0.79. Agreement was highest for Comparative Analysis ($\kappa = 0.84$), lower for Tracking and Retrieval ($\kappa = 0.77$), and lowest for Integrated Reasoning ($\kappa = 0.73$). Disagreements were adjudicated by a senior annotator using the full videos and the evidence-marking guidelines. The final answer key was updated only when adjudication found a genuine ambiguity or annotation error; otherwise the original key was retained.

Evidence annotation checks. For bounded-evidence items, the independent reviewer verified both answer correctness and timestamp coverage. On the overlap audit, interval agreement was measured

after allowing a 5-second tolerance at each boundary. The median temporal IoU between the author interval set and the reviewer interval set was 0.74, with a 90th-percentile absolute boundary deviation of 11.8 seconds. Items whose answer required evidence outside the marked regions were returned for revision. Global-semantic items were audited by checking whether the answer required distributed narrative, causal, or social evidence rather than a single local interval.

D Implementation Protocol

HOLMES is implemented as a deterministic inference controller around a base MLLM. The controller, rather than the model, owns the search state, budget allocation, Dossier updates, and termination rule. Each module produces a schema-constrained record; malformed records are repaired by a single format-only retry, and no unstructured final response can bypass the Dossier. Substituting a different base model changes the local visual judgments, while the slot-filling loop, graph update rules, and verdict gates remain fixed.

We use three risk modes to parameterize visual inspection: STANDARD for ordinary entity/action slots, FINEDETAIL for OCR, small objects, handwritten text, and sub-stride transient cues, and NARRATIVE for chapter-scale slots whose support is distributed across segments. Failed searches produce coverage certificates with deterministic coverage values; absence-dependent hypotheses can cite them only when the computed coverage exceeds the threshold γ. The MLLM is responsible for semantic judgments such as whether a predicate is fine-detail or narrative, whether a local frame shows a pre-state or post-state, and whether two side-by-side events match. The controller is responsible only for schema validation, budget accounting, coverage computation, graph activation, and verdict gates, which it re-checks on the Tribunal's draft so that no model-proposed verdict bypasses them. Exact module prompts, Dossier schema, coverage functions, and default controller settings are provided in Appendix F, Appendix G, Appendix H, Appendix I, and Appendix J.

The final answer is obtained by deterministic graph readout. Since CLoVR-Bench uses multiple-choice questions, ReadOut returns the selected option and its provenance chain. For aggregate-style tasks, the selected count or range option is returned only when the Dossier establishes a separating interval over competing candidates; otherwise the answer is marked as incompletely verified.

E Limitations and Responsible Release

Limitations. CLoVR-Bench focuses on multiple-choice reasoning over publicly available long-form videos, so it does not cover every language, domain, interaction style, or open-ended generation setting. Its global-semantic annotations support narrative and causal evaluation, but they are less temporally localized than bounded evidence regions. HOLMES is training-free and interpretable, but it still depends on the base MLLM's local visual perception and semantic judgments. Its iterative search and side-by-side binding also increase inference cost relative to a single-pass prompt, and coverage certificates should be read as evidence of what was searched rather than a proof that an event is absent. Controlled perturbations (Appendix K.7) show that refinement and constraint audits recover only part of the loss from missing fact cards or corrupted bindings. Human solvers have unrestricted manual access to complete videos, so model–human gaps are not comparisons under matched viewing budgets.

Release and assets. For each video group, the benchmark release contains the QA annotations (questions, answer options, answers, task labels, and per-video evidence timestamps) and the source videos; it is available at `https://huggingface.co/datasets/CLoVR-Bench/CLoVR-Bench`. The videos were collected from publicly available web platforms and remain the property of their original creators. They are provided solely for non-commercial research under a research-only license that prohibits commercial use, redistribution of the videos outside the release, and identification or profiling of individuals; rights holders may request removal of their content. The dataset card documents the file layout and annotation fields. HOLMES prompts, schemas, and controller settings are given in Appendices F–J to support reproduction.

Compute. All reported experiments are inference-only. Open-source MLLMs require GPU workers capable of serving the corresponding public checkpoints at their recommended precision and context length, while closed-source systems are evaluated through their public APIs. We report the per-video

frame budget and HOLMES controller settings in the main paper and Appendix J; measured calls, tokens, latency, and API cost are reported in Appendix K.3. The Qwen2.5-VL-7B measurements use one H100-80GB at batch size one. Appendix K.6 reports scaling with video count and duration; runtime remains dependent on the serving stack and API cost on the dated pricing assumptions in Appendix K.3.

Annotation labor and broader impact. CLoVR-Bench is created by expert annotators following written task definitions, evidence-marking instructions, and quality-control checklists. Annotators are compensated for data curation and annotation labor according to applicable local minimum-wage or institutional requirements, and the work does not collect private participant data or conduct an intervention on human subjects. The benchmark can support safer evaluation of long-video MLLMs by exposing failures in evidence retrieval, cross-video linking, and unsupported answering. Possible misuse includes applying stronger long-video reasoning systems to sensitive or private video collections; we therefore release the benchmark only for non-commercial research under terms that prohibit identifying or profiling individuals, and recommend use only on data for which users have appropriate rights and consent.

F Prompts and Schemas

This appendix specifies the controller-facing prompts and output schemas used by HOLMES. The exact wording can be adapted to a model provider's formatting requirements, but the fields are fixed. MLLM module outputs are parsed as JSON; invalid outputs are repaired once by a format-only retry. The scheduler is not an additional agent module: it is a deterministic controller routine whose records use the same schema discipline.

Component	Input	Output record	Call type
Shared instruction	Module-specific payload	JSON-only grounding and abstention rules	System
Detective	Question, options, dossier summary	Decision slots, hypotheses, evidence slots	MLLM
Scheduler record	Evidence-slot pool, dossier graph, budget	Ready slot instances and estimated cost	Deterministic
Observer	Scheduled slot, frames or segments, references	Fact cards, semantic cards, coverage certificates	MLLM
Cross-Examiner	Dossier updates and pair budget	Binding, redundancy, stance, and revocation edges	MLLM
Tribunal	Deduplicated dossier graph and hypotheses	Draft scores, gates, verdict, missing-evidence report; re-checked by the controller	MLLM (text-only)

F.1 Shared System Instruction

All modules receive the following shared instruction.

Shared System Instruction

```
You are one module in HOLMES, a training-free framework for
cross-long-video reasoning. You must not answer the user
question directly unless you are the Tribunal. Return only
valid JSON matching the requested schema. Ground every
non-neutral claim in a fact card, semantic card, coverage
certificate, or accepted binding edge. If evidence is
insufficient, use "unknown" rather than guessing.
```

F.2 Detective Prompt

The Detective receives the question, answer options, and the active Dossier summary. It emits implementation-level decision slots, hypotheses, and evidence slots.

Detective Prompt: Evidence-Slot Planner

```
Module: Detective.
Input: question, options, active_dossier_summary.
Task:
1. Create decision slots and one hypothesis per answer option.
2. Create typed evidence slots s_k = (p, r, lambda, m, c,
   tau, V, dep, disc).
3. Ensure each search predicate p is option-discriminating but
   answer-neutral.
4. Specify readout attribute r, link variable lambda, risk
   mode m, coverage requirement c, scope, dependencies, and
   discriminators. Assign risk by rubric:
   Standard = ordinary event with reliable state envelope;
   FineDetail = OCR/small object/handwriting/transient cue or
   no persistent aftermath; Narrative = chapter-level causal,
   social, or narrative state. Do not inspect or infer from
   video content.
Return JSON:
{
  "decision_slots": [
    {"slot_id": "...", "slot_type": "choice|aggregate",
     "alternatives": ["..."], "status": "open|resolved"}
  ],
  "hypotheses": [
    {"hypothesis_id": "...", "slot_id": "...", "claim": "..."}
  ],
  "evidence_slots": [
    {
      "evidence_slot_id": "...",
      "decision_slot_id": "...",
      "predicate": "...",
      "readout_attribute": "...",
      "link_variable": "none|...",
      "state_envelope": {"pre": "...", "event": "...", "post": "..."},
      "scope": {"videos": ["..."], "time_range": "unknown|..."},
      "slot_type": "intra|linkage|aggregate|narrative",
      "risk_mode": "Standard|FineDetail|Narrative",
      "risk_rationale": "...",
      "coverage_requirement": {"min_cov": 0.0,
                               "min_duration_sec": "unknown|..."},
      "dependencies": ["fact_card_or_binding_id"],
      "discriminators": [
        {"hypothesis_pair": ["h_i", "h_j"],
         "readout_attribute": "...",
         "separating_condition": "..."}
      ]
    }
  ]
}
```

F.3 Scheduler Record Schema

The scheduler ranks ready slot instances by Eq. 1. It is implemented deterministically by the controller; no MLLM call is used for this step. Its output record is:

Scheduler Record: Deterministic Slot Selection

```
Component: SchedulerRecord.
Input: evidence_slot_pool, active_dossier_graph,
remaining_frame_budget.
Task:
Select ready slot instances. A slot is ready only if all
dependencies are satisfied. Rank by Disc * (1 + Debt) / Cost.
Use risk_mode only to choose localization strategy and coverage
threshold, not to inflate priority. Compute Cost from expected
anchors/segments plus refinement budget.
```

```
Return JSON:
{
  "scheduled_instances": [
    {
      "instance_id": "...",
      "evidence_slot_id": "...",
      "video_id": "...",
      "time_scope": "...",
      "reference_ids": ["..."],
      "risk_mode": "Standard|FineDetail|Narrative",
      "priority_reason": "...",
      "disc": 0,
      "cost": 0,
      "estimated_cost": {"frames": 0, "calls": 0}
    }
  ]
}
```

F.4 Observer Prompt

The Observer fills scheduled evidence slots according to risk mode.

Observer Prompt: Risk-Conditioned Evidence Writer

```
Module: Observer.
Input: scheduled_instance, evidence_slot, sampled_frames_or_segments,
reference_evidence_if_any, remaining_budget.
Task:
If risk_mode is Standard, perform coarse anchor labeling and
between-bracket refinement. If risk_mode is FineDetail, perform
dense scan and crop inspection. If risk_mode is Narrative, write
segment-level semantic cards. Write coverage certificates for
failed searches; never write a negative fact.
Return JSON:
{
  "anchor_state_labels": [
    {"anchor_time": "...",
     "state": "Pre|Event|Post|Ambiguous|Absent",
     "visible_cues": ["..."]}
  ],
  "fact_cards": [
    {"entry_id": "...", "evidence_slot_id": "...", "video_id": "...",
     "interval": ["...", "..."], "readout": {"attribute": "...",
     "value": "..."}, "supporting_frames": ["..."],
     "visual_confidence": 1, "evidentiary_confidence": 1}
  ],
  "semantic_cards": [
    {"entry_id": "...", "evidence_slot_id": "...", "video_id": "...",
     "segment": ["...", "..."], "narrative_state": "...",
     "entity_roster": ["..."], "causal_hooks": ["..."],
     "visual_confidence": 1, "evidentiary_confidence": 1}
  ],
  "coverage_certificates": [
    {"entry_id": "...", "evidence_slot_id": "...", "video_id": "...",
     "searched_scope": "...", "risk_mode": "...", "stride_sec": 0,
     "searched_fraction": 0.0, "temporal_coverage": 0.0,
     "visual_sufficiency": 0.0, "coverage": 0.0,
     "admissible_for_absence": false,
     "recall_ladder_required": true}
  ]
}
```

F.5 Cross-Examiner Prompt

The Cross-Examiner constructs and audits the binding graph.

Cross-Examiner Prompt: Binding-Graph Auditor

```
Module: Cross-Examiner.
Input: active_dossier_graph, new_fact_cards, new_semantic_cards,
new_coverage_certificates, pair_budget.
Task:
1. Apply relation-specific hard feasibility rejection.
2. Score remaining candidates using entity-signature and
   scene-context compatibility as soft signals.
3. Prioritize pairs whose acceptance or rejection could affect
   active hypotheses, then compare top pairs side-by-side.
4. Mark a binding accepted only when decision=accept, visual
   and evidentiary confidences pass threshold, required cue,
   consistency check, and falsifier are present, and no hard
   constraint is violated.
5. Add binding, temporal, causal, redundancy, support, and
   refute edges; mark pruned pairs Unknown.
6. Revoke edges whose falsifiers are triggered and retag stances.
Return JSON:
{
  "binding_edges": [
    {"edge_id": "...", "binding_type":
     "SameEntity|SameEvent|TemporalOrder|NarrativeContinuation|Causal",
     "source_entries": ["..."], "target_entries": ["..."],
     "decision": "accept|reject|unknown",
     "discriminating_cue": "...", "consistency_check": "...",
     "falsifier": "...", "visual_confidence": 1,
     "evidentiary_confidence": 1,
     "accepted": false}
  ],
  "redundancy_groups": [
    {"group_id": "...", "representative": "...", "duplicates": ["..."]}
  ],
  "stance_edges": [
    {"entry_or_edge_id": "...", "hypothesis_id": "...",
     "stance": "Support|Refute|Neutral", "preconditions": ["..."]}
  ],
  "revoked_edges": [{"edge_id": "...", "reason": "..."}]
}
```

F.6 Tribunal Prompt

The Tribunal drafts the record below; the controller re-checks Eq. 3 and G1–G5 (Appendix I) before accepting a verdict.

Tribunal Prompt: Gated Graph Readout

```
Module: Tribunal.
Input: active_deduplicated_dossier_graph, hypotheses, epsilon, gamma.
Task:
Compute support scores. Return an answer only if G1--G5 pass:
margin, positive support, binding validity, constraint consistency,
and no duplicate inflation. Otherwise return a structured
missing-evidence report.
Return JSON:
{
  "scores": [{"hypothesis_id": "...", "score": 0.0}],
  "gates": [{"gate": "G1|G2|G3|G4|G5",
             "status": "pass|fail", "reason": "..."}],
  "verdict": {"status": "resolved|underdetermined",
              "winner": "...", "answer": "...",
              "provenance": ["..."]},
  "missing_evidence": [
    {"slot_id": "...", "competing_hypotheses": ["..."],
     "missing_discriminator": "...",
     "bottleneck": "coverage|visual_ambiguity|linkage|temporal|aggregation"}
  ]
```

```
}
```

G Case Dossier Schema

The Case Dossier $\mathcal{D}$ is stored as a typed graph. Node types are DECISIONSLOT, HYPOTHESIS, EVIDENCESLOT, FACTCARD, SEMANTICCARD, and COVERAGECERTIFICATE. Edge types are SUPPORT, REFUTE, NEUTRAL, BINDING, TEMPORALORDER, CAUSAL, CONSTRAINT, and REDUNDANCY. Fact and semantic cards store source video, temporal scope, originating evidence slot, supporting frame or segment identifiers, and visual/evidentiary confidence. Coverage certificates store searched scope, search trace, risk mode, computed coverage, and whether the certificate is admissible for absence support. Every stance edge stores its preconditions; if a precondition binding is revoked, the edge is deactivated and recomputed before the next Tribunal call.

Formally, an evidence slot is stored as

$$s_k = (p_k, r_k, \lambda_k, m_k, c_k, \tau_k, V_k, \mathrm{dep}_k, \mathrm{disc}_k),$$

where p_k is the option-discriminating but answer-neutral predicate, r_k is the readout attribute, λ_k is an optional link variable, m_k is the risk mode, c_k is the coverage requirement, (τ_k, V_k) is the search scope, and $\mathrm{dep}_k, \mathrm{disc}_k$ store dependencies and discriminators.

H Coverage Functions

Coverage is computed from the search trace, not generated as model self-confidence. For STANDARD slots,

$$\mathrm{Cov}_{std}(\omega) = \min\Big(1,\ \tfrac{|\tau_\omega^{\text{searched}}|}{|\tau_\omega|} \cdot \kappa(\text{stride}) \cdot \eta(\text{state-persistence})\Big), \tag{2}$$

where κ decays when anchor density is below the slot's risk-mode requirement, and η reflects whether the event is expected to leave an observable state-envelope signature. For FINEDETAIL slots, we separate temporal coverage from visual sufficiency:

$$\mathrm{Cov}_{fd}(\omega) = \min\left(\frac{|\tau_\omega^{\text{searched}}|}{|\tau_\omega|} \cdot \kappa_{fd}(\text{stride}, d_{\min}),\ \nu(\text{resolution}, \text{crop}, r_k)\right),$$

where $d_{\min}$ is the slot's minimum-duration assumption and $\kappa_{fd} = 0$ whenever the stride is larger than the duration required for an admissible absence claim. The visual-sufficiency term ν is computed from observable inspection metadata such as whether full resolution or crop inputs were used, whether the target region was visible, and whether the readout attribute r_k is legible at that scale. A full temporal scan with inadequate crops still cannot support an absence claim. For NARRATIVE slots,

$$\mathrm{Cov}_{nar}(\omega) = \frac{\sum_j |\tau_j|\, \mathbf{1}[\text{segment } j \text{ inspected}]}{|\tau_\omega|},$$

optionally multiplied by a scene-cut coverage factor when automatic segmentation fails. Certificates with coverage below γ cannot support absence-dependent hypotheses and instead create coverage debt for the scheduler.

I Scheduling, Localization, and Binding Details

Each discriminator emitted by the Detective is a record (h_i, h_j, r_k, ϕ) saying that readout attribute r_k can separate hypotheses h_i and h_j under condition ϕ. The controller computes

$$\mathrm{Disc}(s_k, \mathcal{H}, \mathcal{D}) = \sum_{(h_i,h_j,r_k,\phi)\in \mathrm{disc}_k} \mathbf{1}[h_i, h_j \text{ are active in } \mathcal{D}]\mathbf{1}[\phi \text{ is unresolved in } \mathcal{D}].$$

Round-zero discrimination is determined by slot metadata emitted from the answer options rather than by an additional controller-side model call.

The scheduling cost in Eq. 1 is a deterministic proxy for visual inspection budget:

$$\mathrm{Cost}(\omega) = A_{m_k}(\tau_\omega) + R_{m_k}.$$

For STANDARD, $A_m(\tau) = \lceil |\tau|/60\,\mathrm{s} \rceil$ and R_m is the between-bracket refinement budget. For FINEDETAIL, $A_m(\tau) = \lceil |\tau|/15\,\mathrm{s} \rceil$ plus the crop budget. For NARRATIVE, $A_m(\tau)$ is the number of scene-cut or fallback segments. We use this only for relative scheduling; reported experiments still account for actual frames and calls.

For standard-mode refinement, state labels are interpreted as follows. Compatible neighboring PRE and POST anchors open a transition interval for densification. AMBIGUOUS anchors or neighbors that do not form a transition open local zoom windows. Adjacent EVENT anchors are merged when their readout and referent are consistent, and kept as separate fact-card candidates otherwise.

A binding edge is accepted when the Cross-Examiner returns ACCEPT, provides non-empty discriminating cue, consistency check, and falsifier fields, has visual confidence at least θ_v and evidentiary confidence at least θ_e, and survives the constraint audit. We set $\theta_v = \theta_e = 3$ on the 1–5 schema scale unless otherwise stated.

Readout uses the support-difference score

$$\sigma(h_k) = \sum_{e \in \mathcal{D}^\dagger} w(e)\big(\mathbf{1}[e \text{ supports } h_k] - \mathbf{1}[e \text{ refutes } h_k]\big). \tag{3}$$

This score controls investigation only; the gates below decide whether a verdict is admissible.

A verdict $\hat{h} = \arg\max_k \sigma(h_k)$ is accepted only if the controller confirms that the following readout gates pass: (G1) $\sigma(\hat{h}) - \max_{k \neq \hat{h}} \sigma(h_k) \geq \epsilon$; (G2) $\hat{h}$ has admissible positive support, with absence-dependent hypotheses requiring a coverage certificate satisfying $\mathrm{Cov} \geq \gamma$; (G3) every binding edge contributing to $\hat{h}$ is accepted and not revoked by the audit; (G4) no active constraint violation involves evidence supporting $\hat{h}$; and (G5) support does not depend on within-redundancy-group duplicates.

J HOLMES Budget Policy

The default policy uses STANDARD for ordinary entity/action slots, FINEDETAIL for OCR, small objects, handwritten text, and sub-stride transient cues, and NARRATIVE for chapter-level evidence slots. The controller may override the Detective-assigned risk mode upward in risk, but not downward. A final absence-supporting certificate must pass at least one recall-ladder step beyond the initial failed search. Unless otherwise stated, we use the following reproducible controller settings. When an experiment specifies an external frame cap, controller proposals are truncated to that cap before Observer execution. The stride orders the proposal pool and does not override the external cap. The parameter K bounds candidate pair comparisons per round, not MLLM calls: multiple visual pair checks can be packed into one call. Actual consumption and packing are reported in Appendix K.3.

Parameter	Default setting
T	3 investigation rounds
ϵ	0.4 Tribunal margin
γ	0.7 minimum coverage for absence support
K	64 side-by-side binding comparisons per round
θ_v, θ_e	3 on the 1–5 binding-confidence scale
b	3 refinement steps per unresolved transition interval
Standard stride	1 frame/minute at full resolution
FineDetail stride	1 frame/15 seconds plus local crops when supported
Narrative segments	scene-cut aligned, fallback 5–10 minute partitions
B	Raw proposal cap $\min(768,\ \sum_{\omega \in \Omega_t} B(\omega))$ frames per round, before equal-frame truncation

FINEDETAIL absence is admissible only when the stride is compatible with the slot's minimum-duration assumption; otherwise the certificate remains below γ and the system either escalates density through the recall ladder or returns UNKNOWN.

K Additional Experiments and Analyses

This section extends the benchmark evaluation with additional backbones, denser visual inputs, measured resource consumption, controlled budgets, and injected upstream errors. Scores are percentages. For CLoVR-Bench, Overall denotes the macro-average over the 14 tasks; subgroup accuracies are item-weighted where explicitly indicated. A shared resource ceiling limits each method's consumption but does not imply identical realized consumption. We distinguish unique source frames from cumulative visual tokens, which also count repeated views and crops.

K.1 Additional Model Evaluations

Table 6 extends evaluation to four additional open models, InternVL3.5-8B [34], LLaVA-OneVision-1.5-8B [1], Qwen3.5-9B [25], and Qwen3.6-35B-A3B [26], on all 2,000 items and Gemini-3.1-Pro Preview [9] on a fixed 500-item subset stratified by task and group size. The subset uses the same 14-task macro-average and is reported separately from full-benchmark results. HOLMES improves every evaluated backbone, while substantial headroom remains. Humans score 89.2 on the same 500-item subset using the protocol in Appendix C, compared with 68.6 for HOLMES. Humans inspect complete videos with manual scrubbing; this 20.6-point gap does not use a matched viewing budget.

Table 6: Additional evaluations on CLoVR-Bench. The last row uses a fixed 500-item subset and should not be ranked against full-benchmark rows.

Model	Items	Direct Overall	+ HOLMES Overall
InternVL3.5-8B	2,000	36.9	48.7
LLaVA-OneVision-1.5-8B	2,000	38.2	49.4
Qwen3.5-9B	2,000	40.5	51.0
Qwen3.6-35B-A3B	2,000	48.7	59.0
Gemini-3.1-Pro Preview	500	60.8	68.6

K.2 Sampling Density, Resolution, and Finding versus Linking

Temporal coverage trades off against spatial detail. Table 7 specifies the resolution and token budgets behind Table 3 and extends the sweep to 256 frames/video. The 720-pixel cap applies to the 32-frame protocol. At 64–256 frames/video, Qwen2.5-VL-7B uses lower resolutions to keep joint visual input near 44.4K tokens/query. Tokens/frame are measured means after image quantization across source aspect ratios, rather than a fixed token count implied by the pre-resize cap.

Table 7: Direct Qwen2.5-VL-7B: frame–resolution trade-off on all 2,000 queries. K denotes thousands of tokens.

Frames/video	Long-side cap	Tokens/frame	Visual tokens/query	Overall
32	720 px	374	37.1K	26.8
64	560 px	224	44.4K	28.5
128	392 px	112	44.4K	29.2
256	280 px	56	44.4K	29.8

At 256 frames/video, a 16:9 frame has a shorter side of approximately 158 pixels. TA, PDR, and ER improve to 27.0, 29.0, and 27.8, respectively (+4.6, +6.5, and +4.3 points relative to 32 frames), while OCR decreases from 39.5 to 34.5. Thus, a larger frame count alone does not describe the information available to the model, and this sweep is not an intervention on temporal sampling at fixed spatial resolution.

Dense and group-level source pools. Table 8 evaluates Qwen3.6-35B-A3B under per-video, per-group, and dense source pools. In the densest setting, sampling is requested at 1 FPS but joint video input is capped at 224K tokens. The realized mean is 612.7 unique source frames/query, or approximately one frame per 15 seconds; this is not realized 1-FPS input.

With the 32-frame/video pool, HOLMES uses 99.2 unique source frames, 61.8K cumulative visual tokens, and 14.2 calls/query. Dense direct inference uses 612.7 frames, 219.6K visual tokens, and one

Table 8: Qwen3.6-35B-A3B on CLoVR-Bench with different source pools. These are system comparisons; cumulative tokens and calls are not matched.

Source pool	Direct Overall	+ HOLMES Overall
32 frames/video	48.7	59.0
256 frames/group	51.5	61.7
1 FPS requested, 224K-token cap	53.4	63.0

call/query. With the dense pool, HOLMES uses 231.5K cumulative visual tokens and 22.6 calls/query because crops and pair checks revisit evidence. Its 63.0 score therefore demonstrates a gain with a dense source pool, rather than an advantage at identical compute. The full-benchmark human score is 88.6 with unrestricted viewing, leaving a 25.6-point gap under different viewing protocols.

The 256-frame group budget allocates an average of 128.0, 74.8, and 47.4 unique source frames/video to groups of two, three–four, and five–six videos, respectively. Direct accuracy in these bins increases from 53.6/47.8/38.7 to 57.0/50.6/41.2. These accuracies are item-weighted within each bin (800/950/250 queries), so they need not aggregate to the 14-task macro-average Overall.

Table 9: Direct Qwen3.6-35B-A3B diagnostics by taxonomy level and evidence sparsity under the source pools in Table 8.

Source pool	L1	L2	L3	Dense	Sparse	Needle-like
32 frames/video	58.2	47.2	41.5	60.8	48.1	37.2
256 frames/group	62.0	51.8	42.6	64.7	51.8	38.3
1 FPS requested, 224K cap	64.2	54.2	44.0	67.8	54.0	39.0

Relative to 32 frames/video, the densest direct input improves L1 and Dense cases by 6.0 and 7.0 points, but L3 and Needle-like cases by only 2.5 and 1.8 points (Table 9). These results are consistent with additional coverage helping localized perception and finding more than higher-order integration. They do not isolate linking from every other source of error: sparse evidence may still be missed. Together with the visual-binding ablation, they indicate that increasing temporal coverage does not by itself remove the cross-video reasoning bottleneck.

K.3 Resource Accounting and Practical Cost

Measurement protocol. The Qwen2.5-VL-7B resource measurements below cover all 2,000 queries on one H100-80GB at batch size one, with resident model weights and indexed video seeking. End-to-end latency includes video I/O, retries, and controller work, and excludes one-time model loading. Calls, tokens, and latency are means per query. Unique source frames/video is the total number of distinct decoded source frames across queries divided by the total number of source-video instances. Proposals are deduplicated, so an external frame cap is an upper bound rather than a quota. Visual-token counts are cumulative across all calls, including repeated views and crops; revisiting a frame increases tokens but not unique source frames. Tables 10 and 11 report method-level and module-level consumption, respectively.

Table 10: Realized resources for every method in Table 4 under a 32-source-frame/video ceiling. Tokens and calls are not matched.

Method	Overall	Frames/video	Visual tokens/query	Calls/query	Seconds/query
AKS	30.5	32.0	37.1K	1.0	19.8
FOCUS	31.8	32.0	37.1K	1.0	21.6
Q-Frame	33.0	32.0	37.1K	1.0	23.9
VideoAgent	34.5	31.8	42.6K	8.0	52.7
LVAgent	36.4	31.6	46.8K	10.9	68.4
HOLMES	41.0	31.4	52.3K	13.6	82.1

The Observer and Cross-Examiner account for 85.7% of latency. The Observer inspects 97.3 unique source frames/query plus 7.6 low-resolution crops (184 visual tokens/crop). After typed blocking, the Cross-Examiner performs 18.4 visual pair checks/query, packing a mean of 4.5 pairs/call into 4.1 calls/query. Thus, the $K = 64$ candidate-comparison cap in Appendix J is not a count of model calls.

Table 11: HOLMES module-level resource use per query with Qwen2.5-VL-7B and the 32-source-frame/video external cap.

Component	Calls	Visual tokens	Text tokens in/out	Seconds
Detective	2.1	0	3.0K / 0.8K	5.7
Observer	6.4	37.8K	6.2K / 1.0K	38.9
Cross-Examiner	4.1	14.5K	7.0K / 0.8K	31.5
Tribunal	1.0	0	1.7K / 0.2K	3.2
Decode/seek/controller	0	0	0 / 0	2.8
Total	13.6	52.3K	17.9K / 2.8K	82.1

Captioning control. Caption-then-Fuse uses the same backbone to caption each video independently (at most 192 generated tokens/video), followed by one text-only fusion call. Table 12 shows that denser captioning explains part of the gap to agent baselines. HOLMES remains 4.1 points above the 128-frame captioning control while using approximately 45% fewer visual tokens, but 1.15 times its latency and more calls. At 128 frames/video, separate video captioning permits about 241 visual tokens/frame, whereas the joint-context direct control in Table 7 uses 112. These are distinct context and resolution settings.

Table 12: Captioning controls and HOLMES on CLoVR-Bench with Qwen2.5-VL-7B. HOLMES uses adaptive source selection within the stated external cap.

System	Frames/video	Overall	Calls/query	Visual tokens/query	Seconds/query
Caption-then-Fuse	32	32.7	4.1	37.1K	43.8
Caption-then-Fuse	128	36.9	4.1	95.7K	71.6
HOLMES	≤ 32	41.0	13.6	52.3K	82.1

API accounting. Table 13 reports Gemini-2.5-Pro usage from API metadata; billed output includes candidate and thinking tokens. Costs use the rates recorded with these experiments in July 2026 for prompts up to 200K tokens: $1.25 per million input tokens and $10 per million output tokens. No caching or grounding calls are used. The 10.5-point gain in Table 2 costs 4.23 times the latency and 3.14 times the API charge. These are dated cost estimates, not a claim about current provider pricing.

Table 13: Gemini-2.5-Pro resource usage and historical API cost per query.

System	Calls	Visual input	Text input	Billed output	Seconds	USD
Direct	1.0	25.6K	1.2K	0.9K	21.9	0.0425
HOLMES	12.8	36.4K	16.8K	6.7K	92.6	0.1335

K.4 Comparisons under Common Resource Ceilings

To separate framework gains from additional resource use, we rerun the three agent methods on Qwen2.5-VL-7B with at most 32 unique source frames/video and 40K cumulative visual tokens/query, varying the call ceiling (Table 14). HOLMES exceeds LVAgent by 3.5 points at eight calls. These are common maxima, not identical realized token or call counts.

At 12 calls/query, the call ceiling does not bind the two baselines. The 40K visual-token cap clips 4.8%, 11.6%, and 37.9% of queries for VideoAgent, LVAgent, and HOLMES, respectively, mainly reducing Cross-Examiner pair checks for HOLMES. Five–six-video groups comprise 12.5% of queries and are affected most strongly; their uncapped demand is approximately 3.6 times that of two-video groups. These differences matter when interpreting a single shared ceiling.

Controlled benchmark difficulty. Table 15 applies the same Qwen2.5-VL-7B backbone and common ceilings to MVU-Eval and CLoVR-Bench. The bare model drops 16.3 points on CLoVR-Bench, and all four structured systems drop 12.0–13.0 points. The harder regime is therefore observed across methods. This contrast does not isolate video duration causally: the benchmarks also differ in their tasks and evidence distributions.

Table 14: CLoVR-Bench Overall under common source-frame, visual-token, and call ceilings. All rows use at most 32 unique source frames/video and 40K cumulative visual tokens/query.

Maximum calls/query	VideoAgent	LVAgent	HOLMES
4	33.1	33.7	35.8
8	34.2	35.3	38.8
12	34.4	36.1	40.6

Table 15: Cross-benchmark comparison under ceilings of 32 unique source frames/video, 40K cumulative visual tokens/query, and eight calls/query. The last column is CLoVR-Bench minus MVU-Eval, in percentage points.

Method	MVU-Eval	CLoVR-Bench	Difference
Bare model	43.1	26.8	-16.3
Caption-then-Fuse	44.7	32.7	-12.0
VideoAgent	46.6	34.2	-12.4
LVAgent	47.3	35.3	-12.0
HOLMES	51.8	38.8	-13.0

K.5 Budget-Controlled Generalization

Table 16 compares the bare model, captioning, and agent methods using Qwen2.5-VL-7B on LongVideoBench validation ($N = 1{,}337$), MLVU development multiple-choice tasks ($N = 2{,}175$), and MVU-Eval ($N = 1{,}824$). Every method is limited to 32 unique source frames *per query*, 20K cumulative visual tokens/query, and eight calls/query. These ceilings are stricter than the *per-video* 32-frame and 40K-token ceilings in Table 15; the two sets of MVU-Eval scores must not be mixed.

Table 16: Transfer under common ceilings of 32 unique source frames/query, 20K visual tokens/query, and eight calls/query. LongVideoBench uses validation accuracy and MLVU uses M-Avg over seven multiple-choice tasks.

Method	LongVideoBench	MLVU	MVU-Eval
Bare model	54.7	70.2	42.8
Caption-then-Fuse	56.0	70.9	44.1
VideoAgent	56.8	71.3	46.3
LVAgent	57.2	71.7	47.0
HOLMES	57.9	72.1	51.4

The HOLMES–LVAgent confidence intervals on LongVideoBench and MLVU are $[-0.6, 2.1]$ and $[-0.4, 1.2]$ points, respectively. Both include zero, so these controls do not establish a significant advantage on single-video benchmarks. On multi-video MVU-Eval, the 4.4-point improvement has a 95% confidence interval of $[2.2, 6.5]$ with $p < 0.001$, using a 10,000-resample paired multiway-cluster bootstrap over source groups and video IDs to account for shared sources. The capped LongVideoBench and MLVU scores (57.9 and 72.1) are lower than the original, resource-unmatched HOLMES results (58.4 and 72.5) in Table 5.

K.6 Scaling with Video Count and Duration

Let V be the number of videos, B_f the external source-frame cap per video, C the number of fact cards, T the number of rounds, and K the cap on binding comparisons per round. With fixed controller caps, Observer source-frame acquisition scales as $O(VB_f)$; at fixed sampling rate without that frame cap, it grows with total video duration. Naive candidate pair construction is $O(C^2)$, while typed blocking prunes candidates and expensive visual verification is bounded by $O(TK)$ pair checks. These bounds count evidence inspections, not exact Transformer FLOPs, and do not imply unconditional linear runtime.

Table 17 shows increasing cost with video count and a smaller increase with duration when source frames are capped. The per-video cap limits visual exposure even as longer inputs incur additional search and control work. Neither this measurement nor the bounded pair-verification policy removes the quadratic worst case of candidate construction.

Table 17: Measured HOLMES scaling with Qwen2.5-VL-7B under the 32-source-frame/video external cap. Duration bins contain exactly three videos/query and 100 task- and sparsity-stratified queries each.

Setting	Queries	Calls/query	Visual tokens/query	Seconds/query
Number of videos per query				
2	800	11.1	34.3K	58.0
3–4	950	14.4	57.3K	87.0
5–6	250	18.8	90.9K	140.0
Duration per video, with three videos per query				
30–45 minutes	100	12.8	49.7K	75.4
45–60 minutes	100	13.4	51.6K	82.3
60+ minutes	100	14.1	53.9K	90.8

K.7 Robustness to Retrieval and Binding Errors

We perturb HOLMES with Qwen2.5-VL-7B under the 32-source-frame/video external cap using five fixed seeds, stratifying sampled perturbations by task and evidence sparsity. For retrieval, the eligible set consists of fact cards overlapping at least one gold support interval; the specified fraction is removed after round-one Observer output. For binding, eligible audited decisions agree with the gold links; sampled positive and negative decisions are flipped proportionally after round-one Cross-Examiner output. Rounds two and three may relocalize evidence or revoke corrupted bindings. Global-semantic support without a gold interval is not part of the retrieval-drop denominator. Table 18 reports the resulting accuracy.

Table 18: Overall accuracy under injected upstream errors. Entries are mean $\pm$ standard deviation over five perturbation seeds; variation reflects the sampled corruption, not independent model training runs.

Injected error	0%	10%	20%	30%
Retrieval-card drops	41.0 ± 0.00	40.1 ± 0.18	38.4 ± 0.27	36.6 ± 0.35
Binding-decision flips	41.0 ± 0.00	39.6 ± 0.22	37.5 ± 0.31	35.1 ± 0.44

At 20% retrieval noise, disabling targeted refinement reduces accuracy from 38.4 ± 0.27 to 35.5 ± 0.32, a 2.9-point difference. At 20% binding noise, bypassing constraint audits, revocation, and hard gates reduces accuracy from 37.5 ± 0.31 to 31.7 ± 0.48, a 5.8-point difference. Binding corruption is more damaging than retrieval-card drops at the tested rates, and the recovery mechanisms only partly repair the damage. These interventions diagnose sensitivity to specified errors; they do not estimate the natural prevalence of retrieval or binding failures. The result complements the summary-only-linking ablation in Figure 5 and makes the dependence on reliable upstream evidence explicit.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three explicit contributions—the CLoVR-Bench benchmark, the systematic evaluation of frontier MLLMs, and the training-free HOLMES framework—all of which are substantiated by the benchmark construction in Section 3 and the experimental results presented later in the paper. Scope and assumptions (cross-long-video setting, training-free inference, hierarchical taxonomy) are also explicitly scoped.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E discusses the main limitations: CLoVR-Bench focuses on multiple-choice QA over public long-form videos, global-semantic items are less temporally localized than bounded evidence regions, HOLMES depends on the base MLLM's local visual judgments, and its iterative search and binding introduce additional inference cost.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important

role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical in nature: it contributes a benchmark (CLoVR-Bench) and a training-free inference framework (HOLMES) and does not claim any formal theorems, lemmas, or closed-form theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducibility information is distributed across dedicated appendices. Appendix B specifies the benchmark construction pipeline, including the small-VLM leakage filter configuration (Qwen3-VL-2B and InternVL3.5-2B, 32 uniformly sampled frames per video, longer side at most 720 pixels) and the role of GPT-5.1 and Gemini-2.5-Pro as distractor-writing assistants. Appendix C reports the human-baseline protocol and annotation reliability. Appendix A provides formal definitions of all 36 subtasks. For HOLMES, Appendix D describes the deterministic controller, Appendix F gives verbatim module prompts and JSON schemas for Detective, Observer, Cross-Examiner, and Tribunal, plus the deterministic scheduler record schema, Appendix G defines the Case Dossier graph, Appendix H gives the coverage functions for Standard, FineDetail, and Narrative slots, Appendix I defines the scheduling cost and binding-acceptance rules, and Appendix J lists all default controller hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The benchmark annotations (questions, answer options, answers, task labels, and per-video evidence timestamps) and the source videos for each video group are publicly released at `https://huggingface.co/datasets/CLoVR-Bench/CLoVR-Bench` under the research-only terms described in Appendix E. HOLMES code is not released; the method is training-free, and its prompts, output schemas, and controller settings are given in full in Appendices F–J so that it can be reimplemented.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: HOLMES is training-free and introduces no optimizer. We instead fully disclose the inference configuration. Appendix J tabulates every controller parameter used in the reported experiments. All hyperparameters were fixed prior to the final evaluation and not tuned on the test items.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The main leaderboard uses deterministic single-run decoding without uncertainty estimates. Appendix K.5 adds confidence intervals for budget-controlled transfer comparisons, including a paired multiway-cluster bootstrap on MVU-Eval, and Appendix K.7 reports means and standard deviations over five perturbation seeds. These analyses do not establish statistical significance for every main-table comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are inference-only. Appendices J and K.3 specify controller settings and measure source frames, cumulative tokens, calls, and latency for Qwen2.5-VL-7B on one H100-80GB at batch size one, including module-level accounting. Appendix K.3 also reports Gemini API usage and cost with explicit historical pricing assumptions; Appendix K.6 reports cost by video count and duration. These timings and prices are specific to the measured serving setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics. The benchmark uses publicly available source videos, which are redistributed only for non-commercial research under a research-only license with removal on request, and is designed for research evaluation. Expert annotators are compensated for data curation and annotation labor, and LLM/VLM assistance is restricted to distractor drafting and shortcut probing rather than unsupervised content generation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E discusses positive impacts, including safer evaluation of long-video MLLMs by exposing evidence retrieval and linking failures, as well as potential misuse on sensitive or private video collections. The appendix also describes the research-only release terms, which prohibit identifying or profiling individuals, and recommends use only on data for which users have appropriate rights and consent.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The release includes source videos collected from public web platforms, which were selected manually by expert annotators rather than scraped automatically (Appendix B). Appendix E describes the safeguards: a research-only license that prohibits commercial use, redistribution of the videos, and identification or profiling of individuals, and removal of content at the request of rights holders.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing benchmarks and model families are cited in the main text and references. The source videos come from public web platforms and remain the property of their creators; Appendix E and the dataset license state that they are provided only for non-commercial research and that users must comply with the terms of the original platforms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces two new assets: (i) the CLoVR-Bench benchmark, comprising 2,000 expert-annotated QA pairs over 400 long-form videos with a three-level hierarchical taxonomy of 14 tasks and 36 subtasks (defined in Section 3 and detailed in

Appendix A), along with temporal-interval and cross-video-link annotations; and (ii) the HOLMES framework, whose module specifications are described in the main text. The data construction pipeline, quality-control procedure, and annotation schema are documented in Appendix B. The public release at `https://huggingface.co/datasets/CLoVR-Bench/CLoVR-Bench` includes a dataset card describing the file layout and annotation fields, together with the research-only license.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The work uses expert annotators for data curation and annotation rather than a crowdworker experiment or an intervention on human subjects. The task definitions given to annotators are provided in Appendix A; Appendix B describes the evidence-marking and quality-control protocol; Appendix C reports annotator training and human-baseline details; and Appendix E states that annotators are compensated according to applicable local minimum-wage or institutional requirements. Screenshots are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work uses paid annotation labor for dataset construction, but it does not run human-subject experiments, collect private participant data, or intervene on annotators beyond ordinary data-labeling work. Annotation labor, compensation, and quality-control procedures are described in Appendix B, Appendix C, and Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs and MLLMs are central, non-trivial components of this work and are explicitly described in the paper. In dataset construction (Section 3), frontier LLMs are used to draft distractors and polish question phrasing under expert verification, and compact open-source VLMs are used as shortcut probes for quality control. In the HOLMES framework (Section 4), an MLLM serves as the shared backend for evidence-slot planning, risk-conditioned localization, typed binding verification, and drafting the graph readout, whose gates are re-checked by the deterministic controller. Model versions, prompting strategies, and evaluation protocols are documented in the methodology sections and appendix.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.