MULTIMODAL-CL-BENCH: A BENCHMARK FOR MULTIMODAL CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Can multimodal large language models (MLLMs) learn new knowledge from multimodal context and apply it downstream, rather than merely perceiving and retrieving information already in their weights? We formalize this capability as multimodal context learning (MCL) and introduce **Multimodal-CL-Bench** to evaluate it. The benchmark contains 344 self-contained multimodal contexts—spanning text, images, and videos—with 1,080 tasks and 7,309 fine-grained rubrics. Rather than grouping cases by surface domain, we organize them by the form of contextual learning required, yielding four categories and 20 subcategories that support diagnostic, capability-level evaluation. Cases are built via expert design and a controlled agent-based synthesis pipeline that minimizes pretraining contamination. Evaluating eight frontier MLLMs uncovers three findings: (1) none of the original eight models exceeds 13% task-level success, with the best reaching only 12.8%; (2) context-tracking failures and modality misgrounding are the most frequent annotated error types; (3) scaling test-time reasoning does not reliably help—Thinking variants do not consistently outperform their base models. These results suggest that the central bottleneck of MCL is stable evidence-state organization across heterogeneous modalities, rather than perception capacity or reasoning length.

1 INTRODUCTION

Multimodal large language models (MLLMs) excel at perceiving and reasoning over inputs already aligned with their pretrained knowledge. But many practical tasks—learning a newly defined visual coding system, following a domain-specific manual with diagrams, or reasoning over observations governed by task-specific rules—require knowledge introduced by the context itself and likely absent from pretraining. Solving them is no longer a matter of recall or perception; it requires the model to learn from the context and act on what it has just learned. This need mirrors the shift from prompt engineering and in-context learning Brown et al. (2020); Dong et al. (2024); Liu et al. (2023) toward context engineering and context learning Mei et al. (2025); Mialon et al. (2023).

CL-BenchDou et al. (2026) recently formalized context learning as acquiring new knowledge from textual context and applying it downstream. We study its multimodal counterpart, multimodal context learning (MCL), in which a model must acquire case-specific knowledge–local rules, symbol mappings, procedures, exceptions, temporal patterns, or decision criteria–from multimodal context and reuse it via cross-modal grounding across text, images, charts, tables, and sequential observations.

Existing evaluations capture MCL only partially. General multimodal benchmarks target perception, VQA, document, chart, or video understanding, but rarely isolate learning from context. Existing long-context benchmarks evaluate retrieval, information integration, and reasoning over extended inputs Shaham et al. (2023); An et al. (2024); Yen et al. (2024); Kuratov et al. (2024); Song et al. (2025), but do not necessarily isolate the acquisition and sequential reuse of case-specific knowledge. Multimodal In-Context Learning benchmarks examine whether interleaved demonstrations help Baldassini et al. (2024); Luo et al. (2024); Chen et al. (2025); Zong et al. (2024), but gains can stem from textual signals, pattern matching, or query-only shortcuts. Despite this growing landscape, no existing benchmark isolates whether MLLMs can learn, ground, and transfer new knowledge from complex multimodal contexts—the precise capability MCL demands.

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

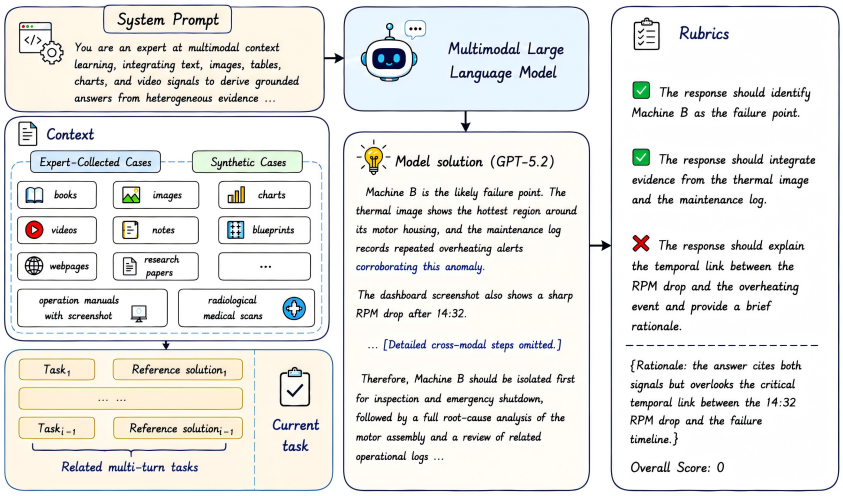

Figure 1: Overview of Multimodal-CL-Bench. Each case provides a system prompt, a multimodal context built from expert-collected and synthetic sources, and a sequence of related tasks. The model must learn task-specific knowledge from heterogeneous evidence, produce a grounded response for the current task, and its response is evaluated by fine-grained task-level rubrics. The illustrated failure shows why partial evidence use is insufficient: a response can identify the correct failure point and cite relevant signals, yet still fail if it misses the required temporal relation.

To address this gap, we introduce **Multimodal-CL-Bench**, a benchmark for evaluating multimodal context learning. As illustrated in Figure 1, each case is *a self-contained multimodal task environment* whose context supplies the evidence, local rules, output constraints, and exception handling required to solve one or more sequential tasks, where later tasks may depend on conclusions established earlier. Crucially, we organize cases not by surface domain (medical, legal, finance, etc.) but by the *form of contextual learning* each task requires—rule induction, temporal state tracking, evidence integration, or standard application—so that benchmark scores translate directly into capability-level diagnostics rather than aggregate performance numbers. Cases are constructed from two complementary sources: expert-collected materials covering specialized or long-tail knowledge, and a controlled agent-based synthesis pipeline that converts traceable sources into closed-world contexts with local rules, distractors, and rubric-annotated tasks. The resulting benchmark contains ***344 contexts***, ***1,080 tasks***, and ***7,309 rubrics***, with inputs spanning images, videos, charts, forms, documents, UI screenshots, and multi-turn histories.

The four categories are: *Rule Learning & Complex Reasoning* (90 contexts), which tests whether models can infer and apply case-specific rule systems from multimodal demonstrations; *Temporal Dynamics & Environmental Interaction* (80 contexts), which tests action, state, and resource tracking across videos and sequential observations; *Evidence Integration & Structured Auditing* (96 contexts), which tests fragmented-evidence collection and conversion into reliable intermediate representations; and *Expert Standards & Scenario Decision-Making* (78 contexts), which tests grounded application of manuals, rubrics, and professional criteria. Across all categories, the benchmark contains 18,740 images and 304 videos, reflecting the visually dense nature and diverse media configurations of the benchmark.

The original experiments on eight frontier MLLMs reveal a consistent gap. First, ***no model exceeds 13% task-level success***: even the strongest, GPT-5.2, reaches only 12.8%, indicating that current MLLMs are far from reliable in MCL. Second, ***exploratory labels highlight contextual tracking and grounding difficulties***: context-tracking errors (34.5%) and modality misgrounding (26.7%) are the largest groups. A breakdown by media configuration sharpens this picture—success drops from 15.1% on single-video cases to 9.1% on multi-video and 7.8% on image+video settings, showing an association between media configuration and performance. Third, ***longer reasoning does not reliably help***: Thinking variants of frontier models do not consistently outperform their base versions, and extended reasoning can produce more elaborate but still misgrounded traces. Together, these findings motivate further investigation of whether models can organize multimodal evidence

into stable intermediate representations, maintain and update contextual states over long horizons, and faithfully bind local rules to visual or structured evidence.

Our contributions are summarized as follows:

- We formulate *multimodal context learning* (MCL) as a benchmarkable capability, where models must acquire task-specific knowledge from multimodal context and reuse it for downstream reasoning and decision-making. This formulation shifts evaluation from passive multimodal perception or retrieval toward context-driven knowledge acquisition, grounding, and transfer.
- We introduce Multimodal-CL-Bench, a comprehensive benchmark for MCL with 344 contexts, 1,080 tasks, and 7,309 rubrics. The benchmark is organized by the form of contextual learning required, covering four categories and 20 subcategories, and is constructed from both expert-collected cases and a controlled agent-based synthesis pipeline to provide diverse, traceable, and realistic multimodal task environments.
- Beyond reporting overall scores, we provide exploratory diagnostic analyses of current MLLMs in MCL. We find that (i) context-tracking and modality-misgrounding labels are the largest observed error groups and that (ii) scaling test-time reasoning—including dedicated Thinking modes—does not reliably resolve these failures. These findings motivate work on stable evidence-state organization, cross-modal grounding, and contextual state maintenance.

2 Related Work

Context Learning. Early approaches mainly relied on Prompt Engineering Sahoo et al. (2024); Schulhoff et al. (2024) and In-Context Learning (ICL) Brown et al. (2020); Dong et al. (2024); Dai et al. (2023); Olsson et al. (2022); Xie et al. (2021); Pan et al. (2023); Min et al. (2022), which guide models via instructions or demonstrations. To extend the information available beyond parametric knowledge, the broader paradigm of context engineering Asai et al. (2024); Schick et al. (2023); Packer et al. (2023) introduces external sources via retrieval augmentation Asai et al. (2024); Jiang et al. (2025), explicit memory Packer et al. (2023); Hu et al. (2025); Zhong et al. (2024), and tool use or environment interaction Yao et al. (2022); Schick et al. (2023). Yet richer context does not imply genuine learning ability, motivating evaluations of learning through sequential problem solving Dou et al. (2025) and context learning Dou et al. (2026), which focuses on acquiring new facts, rules, or relations from context and dynamically applying them at inference.

Since real-world knowledge spans images, text, and videos, this naturally extends to Multimodal Context Learning (MCL), which requires aligning semantics across modalities and integrating multimodal evidence. Related work includes in-context instruction tuning, as in Otter Li et al. (2025), and Multimodal In-Context Learning (MICL) using interleaved image-text demonstrations Jiang et al. (2024); Baldassini et al. (2024); Luo et al. (2024); Chen et al. (2025). Studies of MICL report sensitivity to textual signals and limitations in visual context use Baldassini et al. (2024); Luo et al. (2024); Chen et al. (2025). These observations motivate evaluating whether models acquire and apply case-specific knowledge grounded in multimodal context, rather than assuming that demonstration-based gains establish this capability.

Benchmarks for Multimodal Context Learning. Existing evaluation suites broadly cover reasoning Shi et al. (2024); Cobbe et al. (2021); Lightman et al. (2024); Rein et al. (2023); Dua et al. (2019), general task capability Mialon et al. (2024); Yue et al. (2024); Fu et al. (2025), and agentic ability Yoran et al. (2024); Zhou et al. (2023). Benchmarks also evaluate retrieval and tool use Wei et al. (2025), long-context understanding An et al. (2024); Bai et al. (2024), document QA Kočiskỳ et al. (2018), summarization Huang et al. (2021); Zhong et al. (2021), and long-context counting Song et al. (2025). Related analyses examine sensitivity to information position Liu et al. (2024). These evaluations do not necessarily require acquiring new case-specific knowledge. On the multimodal side, studies identify limitations in visual context use Chen et al. (2025); Baldassini et al. (2024); Luo et al. (2024); VL-ICL Bench Zong et al. (2024) further examines whether queries require contextual examples rather than being answerable from pretrained knowledge alone.

This issue is more pronounced under Context Learning. Recent efforts such as CL-Bench Dou et al. (2026) explicitly evaluate whether models can acquire new knowledge from text, shifting

benchmarks from "knowledge retrieval" to "knowledge acquisition," but remain confined to textual settings. To fill this gap, we introduce a systematic benchmark for Multimodal Context Learning, evaluating cross-modal knowledge acquisition, semantic alignment, knowledge integration and rule induction, and transfer reasoning. Unlike benchmarks centered on static recall or single-turn perception, it better matches realistic scenarios and exposes differences in dynamic learning, complex reasoning, and multimodal generalization, offering a more challenging and diagnostic testbed for next-generation multimodal models.

3 MULTIMODAL-CL-BENCH

We describe the taxonomy underlying our benchmark (§3.1), the two-track construction pipeline that combines expert design with controlled synthesis (§3.2), and the task-level rubric-based evaluation framework (§3.3).

Table 1: Statistics of Multimodal-CL-Bench under the updated taxonomy. Img. Num and Vid. Num denote the total number of images and videos found in the corresponding case directories. Abbreviations: RL & CR: Rule Learning & Complex Reasoning; TD & EI: Temporal Dynamics & Environmental Interaction; EI & SA: Evidence Integration & Structured Auditing; ES & SD: Expert Standards & Scenario Decision-Making.

Context Category	#Contexts	#Tasks	#Rubrics	Tasks/Context		Rubric/Task		#Img. Num	#Vid. Num
				Mean	Max	Mean	Max		
RL & CR	90	274	1,567	3.0	5	5.7	18	2,213	104
TD & EI	80	273	1,798	3.4	6	6.6	20	3,913	147
EI & SA	96	358	3,058	3.7	4	8.5	14	11,367	32
ES & SD	78	175	886	2.2	5	5.1	14	1,247	21
Total	**344**	**1,080**	**7,309**	**3.1**	**6**	**6.8**	**20**	**18,740**	**304**

3.1 MULTIMODAL CONTEXT TAXONOMY

To systematically characterize the capability coverage of Multimodal-CL-Bench, we organize all 344 contexts into a four-way taxonomy according to the core *form of contextual learning* required by each task, rather than by surface domains. This design choice serves a diagnostic purpose: domain-based taxonomies (e.g., medical, legal, finance) confound capability with topic familiarity, whereas grouping by the form of learning required—rule induction, temporal tracking, evidence integration, or standard application—makes benchmark scores directly interpretable as capability-level diagnostics. As shown in Figure 2, the benchmark covers four complementary categories: Rule Learning & Complex Reasoning, Temporal Dynamics & Environmental Interaction, Evidence Integration & Structured Auditing, and Expert Standards & Scenario Decision-Making. The outer ring further illustrates representative subtypes under each category.

Rule Learning & Complex Reasoning. This category contains 90 contexts and focuses on whether models can infer latent rules from the provided context and apply them to new queries. The key challenge is not recognizing a familiar domain, but learning a case-specific rule system from multimodal demonstrations, instructions, or examples. These cases include daily rules, visual patterns, math and code logic, puzzle games, and text logic. They require models to handle distractors, resolve priority conflicts, identify symbolic mappings, and perform multi-step reasoning under locally defined constraints.

Temporal Dynamics & Environmental Interaction. This category contains 80 contexts and evaluates a model's ability to understand temporal evolution and interactive environments. Unlike static recognition tasks, these cases require tracking actions, states, paths, resources, scores, or agent behaviors across videos, sequential images, or multi-turn contexts. Representative subtypes include human motion, game play, navigation, sports actions, and agent behavior. The central difficulty lies in maintaining a coherent state trajectory over time and using observed dynamics to answer downstream questions.

Evidence Integration & Structured Auditing. This is the largest category, with 96 contexts, emphasizing evidence extraction, organization, and reuse. These tasks require models to gather clues

from multiple images, charts, forms, receipts, layouts, or scene records, and then convert fragmented multimodal information into a structured evidence base. The subtypes include work scenes, charts and numbers, forms and receipts, layouts and maps, and evidence ledgers. This category tests whether models can build reliable intermediate representations rather than relying on isolated visual impressions.

Expert Standards & Scenario Decision-Making. This category contains 78 contexts and measures whether models can apply task-specific standards, manuals, rubrics, or professional criteria to concrete scenarios. The goal is not to test memorized domain knowledge, but to evaluate whether a model can ground its decision in the given context. The subtypes include device manuals, repair help, medical grading, assembly steps, and exam scoring. These cases often require following procedural constraints, diagnosing faults, assigning grades, or making safety-aware decisions based strictly on the provided multimodal materials.

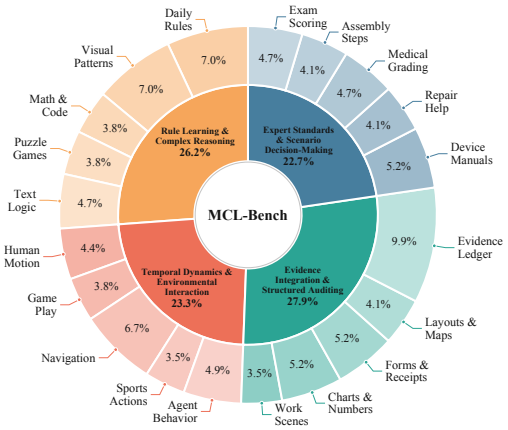

Figure 2: Distribution of context categories in Multimodal-CL-Bench.

3.2 Benchmark Construction

Construction pipeline. Multimodal-CL-Bench is built from two complementary tracks: expert-collected cases and synthetic cases. Expert-collected cases follow three stages—*Case collection*, in which domain experts assemble multimodal contexts spanning text, images, videos, tables, and diagrams that target specialized or long-tail knowledge; *Task design*, in which experts construct sequentially dependent tasks requiring genuine cross-modal evidence integration; and *Rubric design*, in which experts write fine-grained criteria for rigorous task-level evaluation (§3.3). Synthetic cases are produced through a controlled compilation framework detailed below. Both tracks pass quality control to ensure context grounding, multimodal evidence use, and challenge.

Context-grounded design for expert-collected cases. To ensure that Multimodal-CL-Bench evaluates genuine context learning rather than recall from pretraining, we employ three strategies for expert-collected cases to construct contexts containing new knowledge that models are unlikely to have encountered before.

Fictional multimodal contexts. Experts invent entirely new content spanning multiple modalities, such as a fictional legal system with novel case precedents presented through annotated documents, or a new programming language introduced through written specifications and visual execution traces.

Adapted real-world contexts. Experts alter real-world materials to produce new variants, such as revising historical events, changing scientific definitions, or modifying the text and visuals in technical documents, schematics, and charts.

Niche and emerging contexts. Experts include specialized or recently released materials unlikely to be well represented in pretraining corpora, such as new product documentation, recent research with original figures, or professional resources from narrow domains.

These strategies are designed to make task-critical knowledge available through the provided multimodal context rather than through pretrained knowledge alone.

Synthetic case construction. To complement expert-collected cases and expand coverage, we compile traceable multimodal examples into self-contained cases through a four-step pipeline. As shown in Figure 3, these steps form an iterative *Self-Construct* → *Self-Evolve* → *Self-Improve* loop that progressively assembles, refines, and verifies synthetic cases; see Appendix C.1 for details.

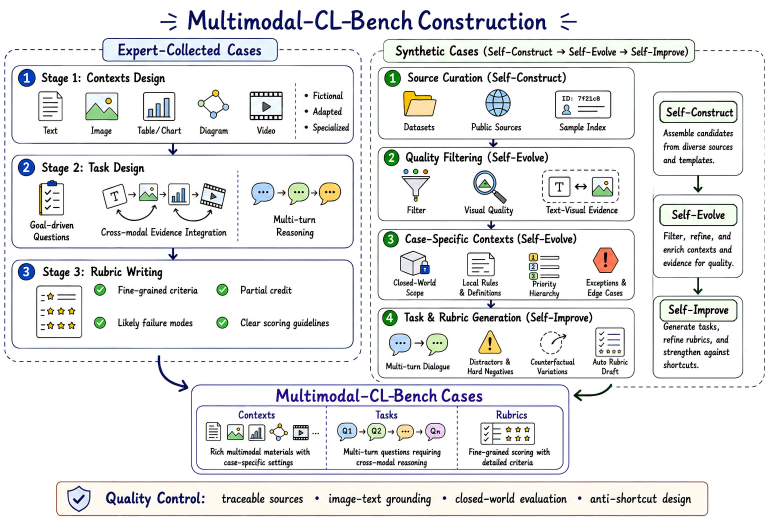

Figure 3: Overview of the Multimodal-CL-Bench construction pipeline. The benchmark consists of two complementary components: expert-collected cases and synthetic cases. Expert-collected cases are built through context design, task design, and rubric writing, while synthetic cases are compiled from traceable multimodal sources through source curation, quality filtering, case-specific context construction, and task/rubric generation. Both components follow quality control to ensure context grounding, multimodal evidence use, closed-world evaluation, and anti-shortcut design.

Source curation. We curate public and author-collected multimodal sources with traceable sample identifiers, accessible assets, and paired visual-textual evidence.

Quality filtering. We filter sources and examples by visual readability, textual support, metadata completeness, traceability, evidence density, and task construction potential, including support for multi-turn reasoning, rule application, counterfactual updates, and rubric-based evaluation.

Case-specific context construction. We transform selected examples into closed-world settings with case-specific labels, rules, priorities, exceptions, and fallback outputs, ensuring that task-relevant knowledge is introduced within the case.

Task and rubric generation. We generate multi-turn tasks and fine-grained rubrics that require models to extract visual evidence, apply case-specific rules, reject distractors, and produce decisions, rankings, corrections, or counterfactual updates. Detailed definitions and examples of these synthetic construction components are provided in Appendix C.

Together, these steps ensure that synthetic cases remain grounded in real multimodal evidence while evaluating multimodal context learning rather than shortcut-based pattern matching or external retrieval.

Validation of construction difficulty. A comparison between the two case sources shows lower task success on synthetic cases: under identical evaluation, average task success drops from **22.42%** on expert-collected cases to **5.43%** on synthetic ones (Appendix C.6). The synthetic track complements expert-collected cases with additional low-pass-rate settings.

3.3 Automatic Evaluation with Task-Level Rubrics

We follow prior work Dou et al. (2026) and use a reasoning-capable LLM as the evaluator, enabling fine-grained, scalable validation of open-ended multimodal responses while remaining aligned with expert judgment.

Our framework is built on high-quality task-level rubrics. For each instance, domain experts decompose the objective into multiple verifiable criteria under a unified protocol, covering factuality, accuracy, instruction adherence, completeness, logical consistency, format compliance, safety, and constraint satisfaction. These rubrics are iteratively refined via cross-review. As shown in Table 1,

Table 2: Performance profile of frontier models on Multimodal-CL-Bench under GPT-5.2 evaluation. Each cell represents the pass rate (%). Abbreviations: RL & CR: Rule Learning & Complex Reasoning; TD & EI: Temporal Dynamics & Environmental Interaction; EI & SA: Evidence Integration & Structured Auditing; ES & SD: Expert Standards & Scenario Decision-Making; Overall: The overall task-level pass rate across all evaluated tasks.

Infer Model	Overall	RL & CR	TD & EI	EI & SA	ES & SD
Claude-Opus-4.6 Thinking	11.6	5.6	12.5	8.3	21.8
Claude-Opus-4.6	11.0	11.1	5.0	6.2	23.1
Claude-Sonnet-4.6 Thinking	11.9	11.1	8.8	8.3	20.5
Claude-Sonnet-4.6	12.5	14.4	6.2	5.2	25.6
Gemini-3.1-Pro-Preview (High)	12.2	13.3	6.2	5.2	25.6
GPT 5.2	12.8	12.2	8.8	9.4	21.8
Doubao-Seed-2.0-pro	8.4	10.0	3.8	2.1	19.2
Gemini-2.5-pro	10.8	6.7	6.2	8.3	23.1

each task contains 6.8 criteria on average, enabling more fine-grained evaluation than single-score benchmarks. We adopt a strict binary all-or-nothing scoring scheme: a task receives 1 only if all criteria are satisfied, and 0 otherwise. GPT-5.2 serves as the primary judge. Further details are provided in Appendix H. Overall, the framework provides a consistent, interpretable, and scalable solution for open-ended LLM evaluation.

4 Evaluation

Our experiments address four questions: (**Q1**) Can current frontier MLLMs reliably learn from multimodal context? (**Q2**) How does this capability vary across different forms of contextual learning? (**Q3**) When models fail, what specifically breaks? (**Q4**) Does scaling test-time reasoning help close the gap?

4.1 Experimental Setup

Models. We evaluate eight frontier MLLMs covering the original eight settings: Gemini-2.5-Pro and Gemini-3.1-Pro-Preview (High); GPT-5.2; Doubao-Seed-2.0-Pro; and four Claude variants—Sonnet-4.6, Sonnet-4.6 Thinking, Opus-4.6, and Opus-4.6 Thinking. The Thinking variants support comparison of reasoning configurations, directly addressing **Q4**; these comparisons do not isolate compute alone.

Evaluation protocol. We adopt a strict task-level rubric-based protocol: a task is counted as solved only if *all* associated rubric criteria are satisfied under binary scoring. GPT-5.2 serves as the primary judge. We report Task Success Rate as our primary metric. Full inference settings, retry policy, and compute accounting are provided in Appendix B.

4.2 Main Results on Multimodal-CL-Bench

(Q1) Context learning in cross-modal environments remains a significant challenge for frontier models. Table 2 shows low task-success rates: **none of the original eight model settings exceeds 13% overall task-level success on Multimodal-CL-Bench**. The strongest model, GPT-5.2, reaches only 12.8%, narrowly leading Claude-Sonnet-4.6 (12.5%) and Gemini-3.1-Pro-Preview (High) (12.2%). The category-level breakdown localizes where the difficulty lies: models perform relatively better on *Expert Standards & Scenario Decision-Making* (best 25.6%), where explicit manuals or rubrics scaffold the reasoning, but collapse on *Temporal Dynamics & Environmental Interaction* and *Evidence Integration & Structured Auditing*, where best results remain below 13% and 10% respectively. This pattern motivates investigation of how models acquire, maintain, and reuse in-situ knowledge grounded in multimodal evidence. On the paired 138-case synthetic subset, ablation studies (Appendix G) further show that removing multimodal evidence while retaining textual rules reduces average success from 6.15% to 2.52%, showing that textual rules alone are insufficient for this subset. A breakdown by media configuration further sharpens this picture: models reach 15.1% on single-video cases but drop to 9.1% on multi-video and 7.8% on image+video settings (Appendix F), showing an observational association with media configuration. Because task diffi-

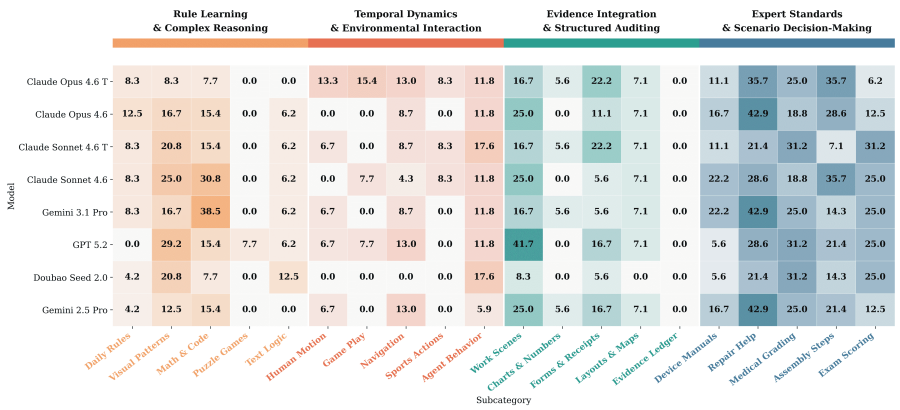

Figure 4: Task-success rates of eight frontier MLLMs across 20 subcategories. Darker cells indicate higher values. No model achieves uniform robustness: high performance in standard-driven subcategories (e.g., Medical Grading, Exam Scoring) does not transfer to subcategories requiring latent state inference (e.g., Navigation, Agent Behavior). Version numbers are omitted for brevity.

culty and case composition are not controlled, these differences do not establish that heterogeneous media cause the decrease or distinguish their effects from context length.

(Q2) Task sensitivity and fine-grained capability gaps. The subcategory heatmap (Figure 4) shows that **no single model is uniformly robust**; instead, performance is highly heterogeneous across the 20 subcategories. Isolated pockets of higher accuracy appear in subcategories with clearer domain cues or explicit evaluation criteria—such as Medical Grading and Exam Scoring—but these gains do not transfer to adjacent subcategories. Performance drops sharply in tasks that require latent state inference or evolving evidence maintenance, including Navigation, Agent Behavior, Form & Receipt, and Evidence Ledger. This descriptive pattern suggests that current models are comparatively more effective when tasks resemble static classification or standard-based judgment, but remain fragile when they must induce rules, preserve intermediate states, or synthesize dense multimodal evidence over extended contexts. The comparison does not isolate causal effects of these task properties.

(Q3) Quantitative analysis of failure modes in multimodal contexts. The exploratory error distribution (Figure 5) highlights **defects in state maintenance and cross-modal grounding**, rather than simple non-responses. *Context Tracking Failure* dominates at 34.5%, indicating that models frequently lose, overwrite, or fail to update intermediate states established earlier in a multimodal interaction. *Modality Misgrounding* accounts for 26.7%, reflecting failures to anchor textual reasoning to the relevant visual entity, region, page, or temporal segment. *Rule/Format Error* (20.0%) and *Context Omission* (18.6%) account for the remainder, while *Refuse to Answer* is nearly absent. Similar error patterns appear across model families. The labels characterize observable defects in model responses rather than their underlying causes.

(Q4) Reasoning configurations show uneven gains on MCL tasks. A natural hypothesis is that longer reasoning chains or dedicated Thinking modes can compensate for the difficulty of MCL. Our results do not show a uniform benefit. As detailed in Appendix D, the Thinking variants of Claude-Opus-4.6 and Claude-Sonnet-4.6 do not consistently outperform their base versions, with Thinking success rates of 11.6% and 11.9%, respectively, and several subcategories even showing lower success under the Thinking configuration. The strongest high-reasoning model, Gemini-3.1-Pro-Preview (High), reaches only 12.2% on average—no better than GPT-5.2 under standard inference. Relatively stronger performance appears in subcategories with explicit textual rules (e.g., Repair Help, where Gemini-2.5-pro reaches 42.9%), but evaporates on tasks requiring temporal ordering or rule induction (Navigation, Game Play, Human Motion all below 10%). A finer exploratory analysis of reasoning *style* (Appendix D) describes associations with reasoning structure—hierarchical audit structuring achieves 19.2% pass rate, while local evidence grounding (the most common style) reaches only 7.5%. This pattern motivates further study of stable evidence-state maintenance across multimodal contexts. Our broader diagnostic analysis across the model development pipeline—covering data curation, pre-training, supervised fine-tuning, and alignment (Appendix J)—discusses targeted recommendations for intermediate structure modeling, cross-modal binding, and long-

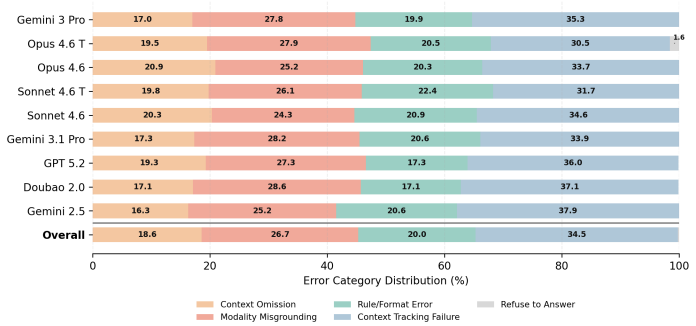

Figure 5: Distribution of context-oriented error types across models under GPT-5.2 evaluation. Failures are dominated by context-tracking errors (34.5%) and modality misgrounding (26.7%), indicating that models often struggle to maintain long-range contextual states or accurately ground answers in visual and structured evidence.

horizon state updating alongside improvements in perception and reasoning. These comparisons are observational: task composition and model configuration may contribute to the differences.

Additional model coverage. Full-benchmark Kimi K2.6 and Kimi K3 results are reported separately in Appendix I.

5 Conclusion

We introduced *multimodal context learning* — the capability to acquire new knowledge from multimodal context and reuse it downstream—and **Multimodal-CL-Bench**, a benchmark of 344 contexts, 1,080 tasks, and 7,309 rubrics organized by the form of contextual learning required. The original evaluation of eight frontier MLLMs yields three findings. First, no model exceeds 13% task-level success, with the best at only 12.8%. Second, exploratory error labels are dominated by context-tracking breakdowns and modality misgrounding. These characterize difficulties with intermediate states and evidence grounding. Third, additional reasoning shows uneven gains—Thinking variants do not consistently outperform their base models, and longer traces often produce more elaborate but still misgrounded answers. These results motivate further work on *stable evidence-state organization*, cross-modal grounding, and contextual state maintenance. We release Multimodal-CL-Bench as a controlled, scalable testbed to drive progress on this axis.

6 AI Use Statement

Generative AI tools were used to assist with academic writing and language editing, the search for and identification of related work, and parts of the research ideation and execution. These tools served only as assistants and did not replace the authors' research judgment, experimental decisions, or final interpretation. All AI-assisted content, references, and research results were manually checked, verified, and revised by the authors.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context

understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1539–1550, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6000–6010. IEEE, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, 2023.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Shihan Dou, Ming Zhang, Chenhao Huang, Jiayi Chen, Feng Chen, Shichun Liu, Yan Liu, Chenxiao Liu, Cheng Zhong, Zongzhang Zhang, et al. Evalearn: quantifying the learning capability and efficiency of llms via sequential problem solving. *arXiv preprint arXiv:2506.02672*, 2025.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. Cl-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In *Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: Human language technologies*, pp. 1419–1436, 2021.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7064–7074, 2025.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557, 2025. doi: 10.1109/TPAMI.2025.3571946.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5321–5335, 2024.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. transformer circuits thread, 2022, 2022.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*, 2024.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, 2023.

Quan Shi, Michael Tang, Karthik Narasimhan, and Shunyu Yao. Can language models solve olympiad programming? *arXiv preprint arXiv:2404.10952*, 2024.

Mingyang Song, Mao Zheng, and Xuan Luo. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3753–3763, 2025.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8938–8968, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, 2021.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. Vl-icl bench: The devil in the details of multimodal in-context learning. *arXiv preprint arXiv:2403.13164*, 2024.

APPENDIX

A Limitations **14**

B Experimental Setup and Reproducibility Details **14**

C Synthetic Case Construction Details **15**

C.1 Iterative Self-Construct, Self-Evolve, and Self-Improve Loop 15

C.2 Source Curation and Self-Construct 15

C.3 Quality Filtering and Self-Evolve 15

C.4 Task/Rubric Generation and Self-Improve 15

C.5 Context-Dependence Checks 16

C.6 Expert-Collected and Synthetic Case Comparison 16

D Impact of Reasoning on Multimodal Context Learning **17**

E Task Success by Number of Images **20**

F Pass-rate Comparison Under Different Modalities **20**

G Effect of Removing Multimodal Evidence **21**

H System Prompt for the LM-Based Verifier **23**

I Supplementary Evaluation of Open-Weight Models **24**

J Discussion **25**

K Failure Case Analysis **26**

K.1 Refuse to Answer 26

K.2 Rule/Format Error 27

K.3 Context Tracking Failure 29

K.4 Context Omission 31

K.5 Modality Misgrounding 34

A LIMITATIONS

Interleaved image-video contexts. Although Multimodal-CL-Bench covers a wide range of multimodal inputs, including images, videos, documents, charts, forms, UI screenshots, and multi-turn histories, cases that tightly interleave static images and dynamic videos remain relatively underrepresented. Such settings are important in real-world multimodal context learning, where a model may need to learn rules or reference structures from static diagrams, screenshots, or documents, and then apply them to dynamic states observed in videos. Expanding the proportion of image-video interleaved cases would allow future versions of Multimodal-CL-Bench to more comprehensively evaluate cross-modal state tracking, evidence alignment, and knowledge transfer across different temporal granularities.

Video sampling constraints. Our evaluation of video-based context learning is constrained by the cost and latency of frontier multimodal API models. To make large-scale evaluation feasible across many cases, tasks, and model families, we cannot sample an arbitrarily large number of frames from each video. As a result, the current benchmark may be less sensitive to extremely fine-grained actions, subtle state transitions, short-lived visual events, or evidence that appears only between sampled frames. The reported results should therefore be interpreted as measuring multimodal context learning under practical API and input-budget constraints, rather than as a complete assessment of dense frame-level video understanding.

Automatic evaluation. Multimodal-CL-Bench relies on rubric-based automatic judging to support scalable evaluation of open-ended multimodal responses. Although the rubrics are designed to be strict and interpretable, and selected cases are checked through auxiliary reviewers and human audits, model-as-a-judge evaluation may still be affected by rubric wording, evaluator behavior, and ambiguous edge cases. Future work could expand human expert auditing, improve evaluator calibration, and explore more cost-effective evaluation pipelines that support denser video sampling and broader image-video mixed scenarios.

B EXPERIMENTAL SETUP AND REPRODUCIBILITY DETAILS

Test details. Table 3 summarizes the key test-time settings shared across all evaluated models.

Table 3: Test-time settings.

Item	Setting
Temperature	1.0
Top-p	0.95
Claude Sonnet Thinking budget	10,000 tokens
Claude Opus Thinking budget	12,000 tokens
Gemini High setting	`thinkingLevel=HIGH`
Maximum retries	20
Request timeout	1800 seconds

Computer resources. Table 4 details the compute setup and API-call accounting for the main reported run.

Table 4: Compute resources and API-call accounting for the main reported run.

Item	Value
Local compute use	CPU
Cases	344
Tasks	1,080
Rubrics	7,309
Case workers	4
Inference workers	50
Evaluation workers	50
Reported inference calls	$1080 \times 8 = 8{,}640$
Reported judge calls	$1080 \times 8 = 8{,}640$
Minimum successful API calls	17,280 before retries

C Synthetic Case Construction Details

This appendix provides additional details on the synthetic case construction process in Multimodal-CL-Bench. The goal of synthetic construction is to complement expert-collected cases by compiling traceable multimodal examples into closed-world cases with case-specific rules, multi-turn tasks, and fine-grained rubrics. Unlike unconstrained generation, the pipeline remains grounded in real multimodal evidence and uses explicit filtering, refinement, and verification steps.

C.1 Iterative Self-Construct, Self-Evolve, and Self-Improve Loop

The synthetic construction pipeline follows an iterative *Self-Construct* $\rightarrow$ *Self-Evolve* $\rightarrow$ *Self-Improve* loop. In *Self-Construct*, the pipeline assembles candidate cases from curated multimodal sources and fixed templates. In *Self-Evolve*, it filters and refines candidates into closed-world settings with local rules, priorities, exceptions, and fallback states. In *Self-Improve*, it generates and revises tasks, rubrics, distractors, counterfactual variants, and anti-shortcut checks. Cases that fail quality or context-dependence checks are revised through earlier stages before being accepted.

C.2 Source Curation and Self-Construct

We curate candidate sources from publicly available multimodal datasets and datasets collected by the authors. A source is eligible only if it provides traceable sample identifiers, accessible assets, and paired visual-textual evidence, such as document images with OCR, charts with questions and answers, diagrams with labels, UI screenshots with descriptions, or visual scenes with annotations. We do not directly use dynamic webpages as benchmark inputs; final assets are materialized locally and linked to source- and sample-level metadata, including dataset name, sample identifier, source URL, license or access status, and local asset path.

During self-construction, selected examples are assembled with a fixed case template. The template constrains the available multimodal evidence, case-specific definitions, output requirements, task progression, rubric coverage, and validity checks. It is used to compile grounded cases from evidence rather than to freely generate answers.

C.3 Quality Filtering and Self-Evolve

Candidate examples are filtered according to whether they can support high-quality multimodal context-learning cases. The filtering process checks whether the visual evidence is readable, the textual or metadata support is sufficient, the sample is traceable, the evidence is dense enough for nontrivial reasoning, and the example can support case-specific rules, distractor rejection, counterfactual updates, and rubric-based evaluation.

After filtering, selected examples are evolved into closed-world case settings. This step introduces a case-specific rule layer on top of the original multimodal evidence. The rule layer defines local labels, evidence-to-label mappings, priorities, tie-breaking rules, exception conditions, and fallback outputs for missing, ambiguous, contradictory, or uncovered evidence. These constraints reduce ambiguity and ensure that the model must use the provided case context rather than infer an answer from external knowledge or a single salient cue.

C.4 Task/Rubric Generation and Self-Improve

For each evolved context, we generate multi-turn tasks and task-level rubrics. Tasks are designed to require models to extract visual evidence, apply case-specific textual rules, reject distractors, update conclusions under local counterfactual changes, and produce structured rankings or decisions. Later turns may reuse intermediate states from earlier turns, such as extracted evidence, local codes, priorities, groupings, corrections, or baseline decisions.

Rubrics are written as fine-grained binary criteria. They check visual evidence extraction, textual rule application, intermediate state tracking, final-answer correctness, distractor rejection, exception handling, counterfactual consistency, output format, and avoidance of external knowledge. A task is considered solved only when all associated rubric criteria are satisfied. During self-improvement,

weak tasks or rubrics are revised to strengthen evidence grounding, remove answer leakage, and improve coverage of likely failure modes.

C.5 Context-Dependence Checks

We apply context-dependence checks to reduce shortcut solutions. A synthetic case is considered valid only if the task cannot be reliably solved after removing key parts of the context. Specifically, we check the following failure conditions:

- **Image ablation**: removing the visual input should make the task substantially harder or unsolvable.
- **Rule ablation**: removing the case-specific textual rules should make the task substantially harder or unsolvable.
- **Context-free ablation**: solving the task without the case context should not be reliable.
- **Shortcut check**: the final answer should not be directly recoverable from a source summary, original QA answer, salient single cue, or real-world common sense.
- **Neighbor-rule check**: when similar rules or labels exist, rubrics should verify why the model does not select a neighboring incorrect rule.

These checks help ensure that synthetic cases evaluate multimodal context learning rather than shortcut-based pattern matching.

C.6 Expert-Collected and Synthetic Case Comparison

We compare expert-collected cases with synthetic cases to examine whether the synthetic construction pipeline can produce useful benchmark instances at scale. The comparison uses **GPT-5.2** as the primary evaluator and reports task-level success from the binary `score` field. We use eight inference models in the comparison. The expert-collected subset contains **155** cases and **325** task instances per model, while the synthetic subset contains **189** cases and **755** task instances per model.

Figure 6 shows a consistent gap between expert-collected and synthetic cases. Across the eight models, the average task success rate drops from **22.42%** on expert-collected cases to **5.43%** on synthetic cases. This decrease appears for every model: the synthetic subset reduces task success by **15.50** to **21.44** percentage points relative to the expert-collected subset. The strongest synthetic-case results are still low, with Claude-Sonnet-4.6 Thinking and Claude-Sonnet-4.6 both reaching only **8.21%**, while weaker models such as Doubao-Seed-2.0-pro and Gemini-2.5-pro reach only **2.65%**.

These results support the feasibility of synthetic construction as a scalable source of difficult multimodal context-learning cases. Importantly, the comparison should not be interpreted as showing that synthetic cases fully replace expert-collected cases in distributional coverage or domain authenticity. Rather, the synthetic pipeline provides a controllable way to generate low-pass-rate stress tests grounded in real multimodal evidence, complementing expert-collected cases with larger-scale and difficulty-oriented coverage.



Figure 6: Task-level success comparison between expert-collected and synthetic cases. Each panel reports one model, with blue-gray bars denoting expert-collected cases and salmon bars denoting synthetic cases. All results use GPT-5.2 as the primary evaluator. Synthetic cases show substantially lower task-level success across all eight models.

D Impact of Reasoning on Multimodal Context Learning

We conduct a systematic comparison between the low-reasoning group Figure 7 and the high-reasoning group Figure 8 across 20 subcategories to examine the relationship between reasoning intensity and multimodal context learning efficiency. The results suggest that increasing the inference budget alone yields only limited gains, and that these gains are unevenly distributed across subcategories. Under the low- or no-reasoning setting, GPT-5.2 provides the strongest baseline performance, with an average score of 12.8%, and shows particularly strong results on evidence-auditing tasks, for example reaching 41.7% in Work Scenes. Under the high-reasoning setting, even the strongest model, Gemini-3.1-Pro-Preview (High), attains an average success rate of only about 12.2%. In addition, the Thinking variants of the Claude series do not show consistent step-level improvements over their standard counterparts, with task-success rates of 11.6% for Claude-Opus-4.6 Thinking and 11.9% for Claude-Sonnet-4.6 Thinking. In several subcategories, their performance also fluctuates rather than improving monotonically. These findings motivate studying how models organize contextual knowledge and apply local rules under long-horizon multimodal constraints.

A further subcategory-level analysis reveals substantial polarization in the cognitive profiles of current frontier models. Temporal Dynamics is consistently among the most challenging regimes across model families. Both baseline and reasoning-augmented models show weak performance on subcategories such as Navigation, Game Play, and Human Motion, with results that are generally low and clearly below those of evidence-auditing and expert-decision tasks. These tasks require models to maintain temporal order and dynamic state transitions across heterogeneous multimodal evidence, which remains difficult even when additional test-time computation is provided.

By contrast, models exhibit relatively stronger performance in Evidence Auditing and Expert Decision categories. For example, Gemini-2.5-pro achieves 42.9% in the Repair Help subcategory, suggesting that current models are comparatively more capable of extracting static multimodal evidence and mapping it to known textual rules in semi-structured expert settings. However, performance still

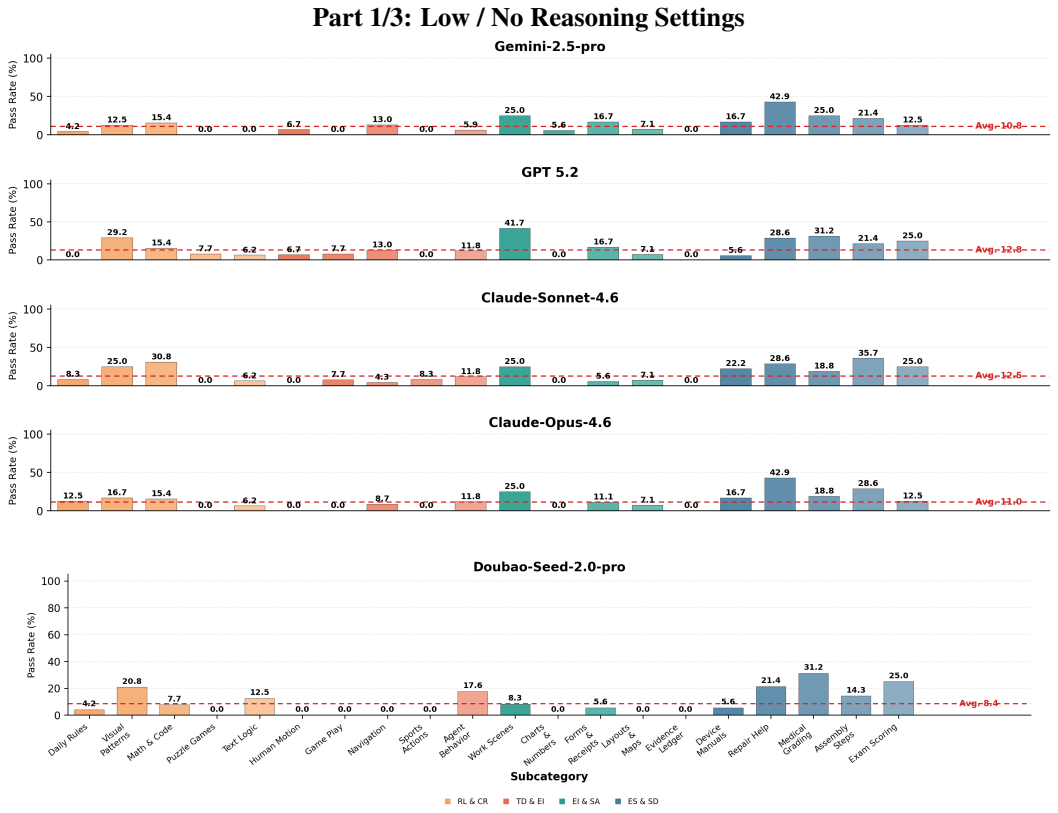

Figure 7: Subcategory performance under Low / No Reasoning settings. This part represents standard zero-shot or direct-response modes. Abbreviations: RL & CR: Rule Learning & Complex Reasoning; TD & EI: Temporal Dynamics & Environmental Interaction; EI & SA: Evidence Integration & Structured Auditing; ES & SD: Expert Standards & Scenario Decision-Making;

degrades substantially on tasks involving complex constraint satisfaction or rule induction from raw evidence, indicating persistent inductive fragility.

Overall, the evaluated reasoning configurations show uneven gains, with low success persisting in several subcategories. The difficulty of Multimodal-CL-Bench suggests that models must move beyond instruction following and develop stronger capabilities for acquiring, organizing, and applying contextual knowledge. The lack of such capabilities remains a central obstacle to more general multimodal intelligence.

Reasoning-style analysis. Figure 9 shows that, at the aggregate level, Local Evidence Grounding and Procedural Constraint Decomposition are the two most prevalent reasoning styles, whereas Hierarchical Audit Structuring and Global Schema Induction appear less frequently. However, their associated pass rates follow a different pattern: Hierarchical Audit Structuring yields the highest success rate (19.2%), followed by Procedural Constraint Decomposition (13.4%), while Local Evidence Grounding has the lowest success rate (7.5%). This suggests that current MLLMs are often able to identify local multimodal evidence, but are less reliable at transforming such evidence into stable and well-structured reasoning processes.

The model-level distributions further indicate that different systems exhibit different reasoning preferences and failure modes. Claude Thinking variants are more frequently associated with Hierarchical Audit Structuring, suggesting that increased reasoning effort in these models often manifests as reorganizing heterogeneous inputs into layered, audit-like representations. This pattern is notable because the potential benefit does not appear to come merely from generating longer responses, but from imposing a more explicit organizational structure over noisy multimodal evidence. GPT-5.2, by contrast, is dominated by Procedural Constraint Decomposition (52.3%), suggesting that it more often frames multimodal tasks as ordered constraint-solving problems involving steps, conditions, and state transitions. Gemini-3.1-Pro-Preview (High) and Gemini-2.5-pro place greater emphasis

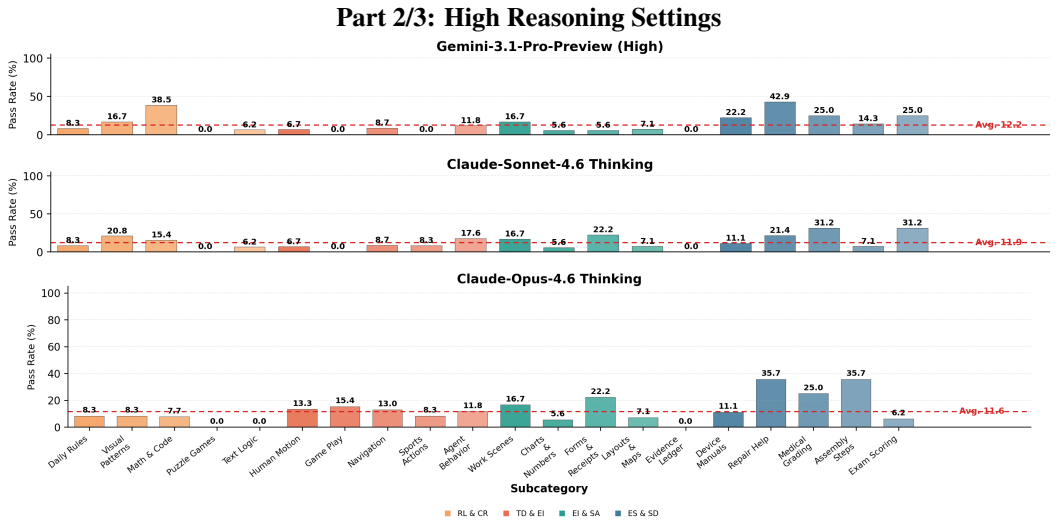

Figure 8: Subcategory performance under High Reasoning settings. This part evaluates models with advanced reasoning features enabled, specifically utilizing Gemini-3.1-Pro-Preview in its High effort mode and Claude-4.6 models in their Thinking (Long-form reasoning) configurations. Abbreviations: RL & CR: Rule Learning & Complex Reasoning; TD & EI: Temporal Dynamics & Environmental Interaction; EI & SA: Evidence Integration & Structured Auditing; ES & SD: Expert Standards & Scenario Decision-Making;

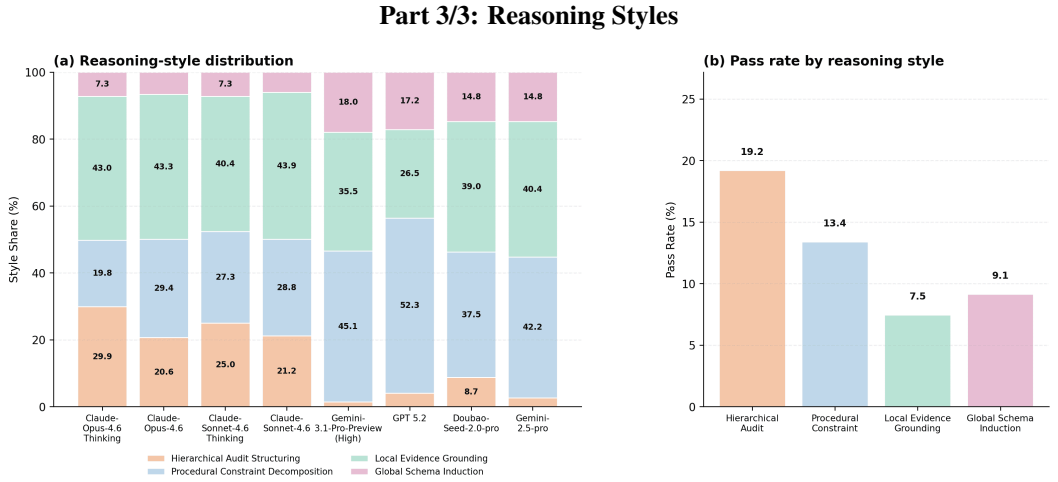

Figure 9: Qualitative reasoning-style analysis of high-reasoning models under GPT-5.2 evaluation. (a) shows the distribution of four reasoning styles inferred from model outputs: Hierarchical Audit Structuring, Procedural Constraint Decomposition, Local Evidence Grounding, and Global Schema Induction. (b) reports the pass rate associated with each reasoning style.

on Procedural Constraint Decomposition and Local Evidence Grounding, which suggests stronger tendencies toward local evidence extraction and rule unfolding, but comparatively less reliance on stable higher-level abstraction. This difference may help explain why some models appear detailed yet remain brittle: they can enumerate visual anchors or procedural fragments, but still fail to maintain correct evidence-to-decision binding across the full context. A representative example is shown in Appendix K, Figure 22, where the model produces a highly organized answer surface while the underlying visual grounding remains misaligned with the task semantics.

The association between reasoning style and success motivates studying whether structured intermediate representations support cross-modal integration and multi-step reasoning. Model and task differences may also contribute to the observed pattern.

E Task Success by Number of Images

We further analyze how task-level success changes with the number of images in pure-photo cases. A case is included only if every task in the case contains at least one image, contains no video, and uses only text and image inputs. This filtering yields **292** pure-photo cases and **952** task instances for each Gemini model. All results use **GPT-5.2** as the primary evaluator and compute success from the binary task-level `score` field.

For each task t, we define the image count as

$$N_{\text{image}}(t) = \sum_{m \in \mathcal{M}_t} \sum_{c \in m} \mathbf{1}[c.\text{type} = \text{image_url}] \,.$$

Here, $\mathcal{M}_t$ is the ordered set of messages available to task t, m denotes one message in that set, and c denotes one content item in message m. Thus, the expression counts all image content items available to task t. We group tasks into six image-count buckets: 1, 2–4, 5–9, 10, 11–30, and 31+.

Figure 10 shows that image count alone does not induce a monotonic difficulty curve. Moderate image counts can improve task success because the context may provide more explicit visual evidence, while larger image sets introduce evidence-selection and cross-image binding burdens. The 10-image bucket is notably low for both models, suggesting that this bucket is dominated by a task template or case family with higher intrinsic difficulty rather than by image count alone. At the aggregate level, the highest average success rate appears in the **5–9** bucket (**22.3%**), whereas the lowest appears in the **10** bucket (**3.4%**). Gemini-3.1-Pro-Preview (High) is stronger than Gemini-2.5-pro in most buckets, with the largest gap in the 5–9 image bucket (**27.7%** vs. **16.9%**).

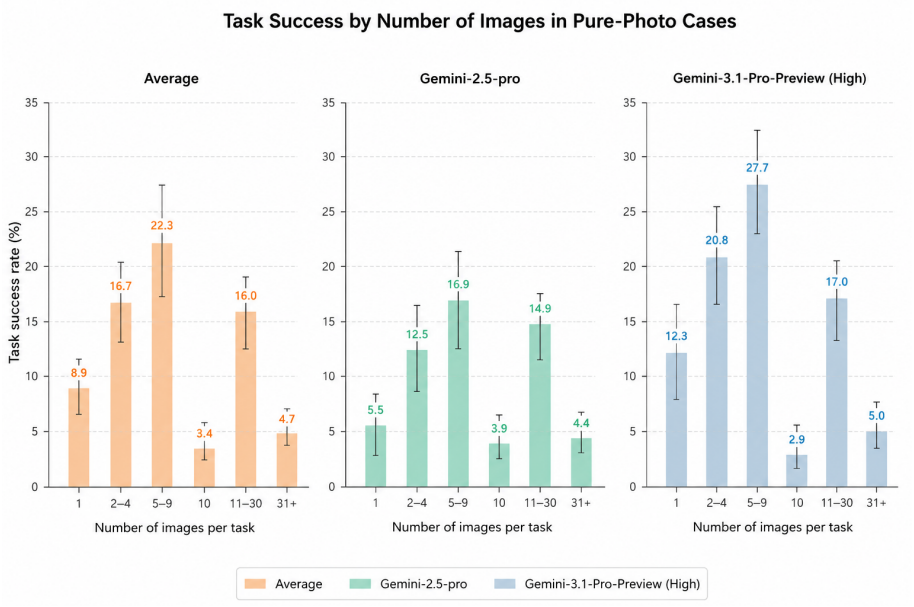

Figure 10: Task success rate by number of images in pure-photo Multimodal-CL-Bench cases. Only case-level pure-photo cases are included: every task in the case must contain at least one image, contain no video, and use only text and image inputs. All results use GPT-5.2 as the primary evaluator and report task-level success from the binary `score` field. Error bars show binomial standard error.

F Pass-rate Comparison Under Different Modalities

The media-configuration groups show different average success rates: Single Video 15.1%, Multi Video 9.1%, and Image + Video 7.8% (Figure 11). These groups differ in case composition and task difficulty. The comparison therefore describes an association with media configuration, not a causal effect of heterogeneous inputs.

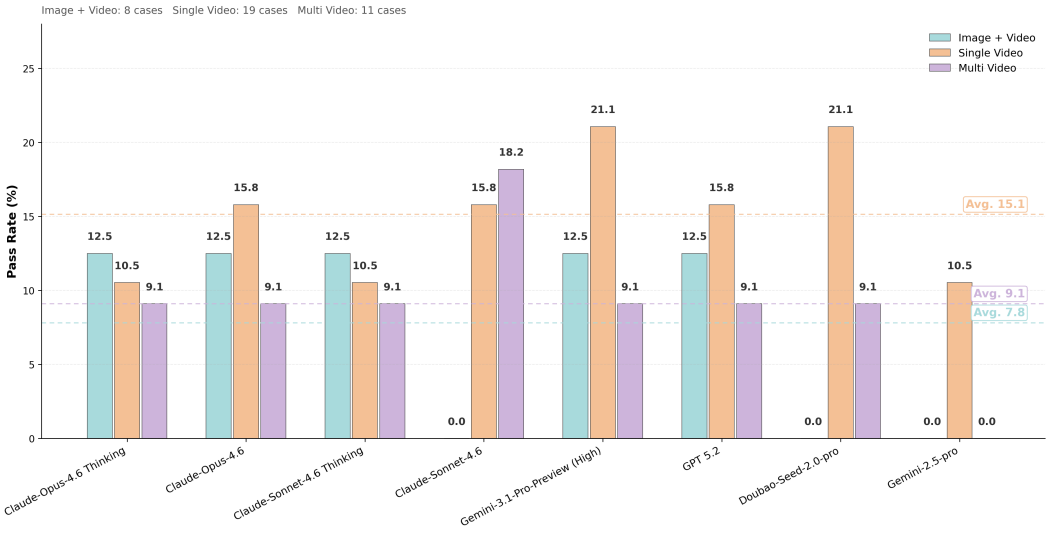

Figure 11: Pass rates by media configuration. Differences are descriptive and may be confounded by case composition and task difficulty.

Image and video evidence can require different grounding and temporal operations, but the present comparison does not isolate these mechanisms from input volume, context length, or task-family differences.

On the other hand, although Multi Video slightly outperforms Image + Video, it remains substantially below Single Video, with most models clustered at low performance levels. Cross-segment state tracking is one possible challenge in multi-video tasks. Models must establish consistent event correspondence, rule abstraction, and environment continuity across multiple video clips, effectively requiring cross-temporal memory and inter-segment consistency modeling. The lower multi-video scores motivate further investigation of how models integrate evidence across clips.

In terms of model differences, sensitivity to multimodal structure varies significantly. Excluding Gemini-3-Pro-Preview (High), the best-performing models still reach around 21.1% on Single Video, but drop to 0%–12.5% on Image + Video and Multi Video. These differences may reflect demands on cross-modal evidence binding and context maintenance, alongside differences in task composition.

These observations motivate controlled comparisons of media configuration.

G Effect of Removing Multimodal Evidence

We further test whether Multimodal-CL-Bench can be solved from textual rules or template cues alone by comparing each case under two input settings: the original setting with multimodal evidence and an ablated setting where the multimodal evidence is removed. For each ablated example, we compare it with the corresponding full-input example using the same case, inference model, and primary evaluator. All results use **GPT-5.2** as the primary evaluator and report task-level success from the binary `score` field. The paired subset contains **138** cases from two synthetic case sources, with **551** paired task instances per model and **4,408** paired model-task evaluations in total.

Figure 12 and Table 5 show that removing multimodal evidence consistently reduces performance. Across the eight models, the average task success rate drops from **6.15%** with multimodal data to **2.52%** without multimodal data, a decrease of **3.63** percentage points. The drop appears for every model, ranging from **2.00** points for Gemini-3.1-Pro-Preview (High) and Doubao-Seed-2.0-pro to **6.90** points for Claude-Sonnet-4.6 Thinking.

These results indicate that the benchmark is not primarily solved through textual rules, memorized templates, or source-level shortcuts.Instead, models benefit from the provided multimodal evidence, and removing that evidence damages task-level success even when the same textual case structure and evaluation protocol are retained.At the same time, the absolute success rate remains low even

with multimodal inputs, showing that visual evidence is necessary but not sufficient: models must still bind evidence to case-local rules, maintain intermediate state, and execute multi-turn reasoning under strict task-level rubrics.

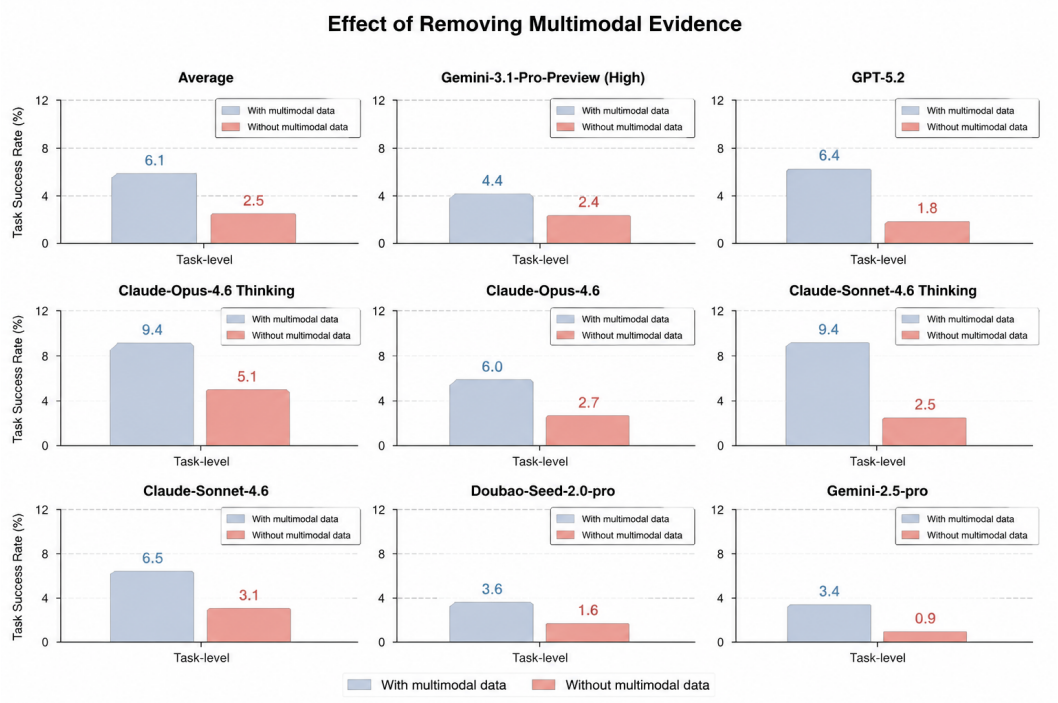

Figure 12: Task-level success with and without multimodal evidence. Each panel reports one model on the same paired cases, with blue-gray bars denoting the original setting with multimodal data and salmon bars denoting the ablated setting without multimodal data. All results use GPT-5.2 as the primary evaluator. Removing multimodal evidence consistently lowers task-level success across all eight models, indicating that Multimodal-CL-Bench cannot be solved reliably from textual rules or template cues alone.

Table 5: Task-level success with and without multimodal evidence on the paired subset. Each model is evaluated on the same 138 cases and 551 task instances.

Model	With MM (%)	Without MM (%)	Drop	Paired Tasks
Gemini-3.1-Pro-Preview (High)	4.36	2.36	-2.00	551
GPT-5.2	6.35	1.81	-4.54	551
Claude-Opus-4.6 Thinking	9.44	5.08	-4.36	551
Claude-Opus-4.6	5.99	2.72	-3.27	551
Claude-Sonnet-4.6 Thinking	9.44	2.54	-6.90	551
Claude-Sonnet-4.6	6.53	3.09	-3.45	551
Doubao-Seed-2.0-pro	3.63	1.63	-2.00	551
Gemini-2.5-pro	3.45	0.91	-2.54	551
Average	6.15	2.52	-3.63	551

H System Prompt for the LM-Based Verifier

We present the system prompt used by the LM-based verifier to grade model solutions in Table 6.

Table 6: System prompt used by the LM-based verifier to enforce strict rubric adherence.

System Prompt: LM-based Verifier

Grading criteria (binary). This is a strict, all-or-nothing grading system. The final score is binary.
Score = 1: Every rubric requirement is satisfied under the rules below.
Score = 0: At least one explicit core requirement is not satisfied.
Constraint Policy: Hard constraints must be enforced strictly only when the rubric clearly says so (e.g., "must", "exactly", "no extra text"); otherwise, evaluate by semantic equivalence.

Grading process. Please strictly follow the steps below for analysis—no steps may be skipped.
Step 1: Analyze the Rubrics and Constraint Types.
For each, decide if it is a hard constraint based on explicit signals (e.g., verbatim, JSON schema must be, must include section number).
Step 2: Check Each Requirement Against the Student's Answer.
Verify requirements one by one under the appropriate constraint type.
Apply Normalization Rules for case, numeric formatting, and whitespace unless the rubric explicitly tests surface form.
Accept semantic equivalence (synonyms, reordered explanations) for non-hard requirements if the core meaning is preserved.
Step 3: Self-Reflection.
Completeness Check: all requirements reviewed with no omissions.
Constraint Check: hard constraints enforced only when explicitly declared.
Semantic Equivalence Check: ensure non-hard form differences were normalized.
Strictness Check: explicit core semantic/content requirements were still evaluated strictly.
Consistency Check: rationale aligns logically with the final score.
Objectivity Check: judgments based on objective facts rather than speculation.

Output format requirements.
[Grading Rationale]: xxx
[List of Requirement Satisfaction Status]: $[x_1, x_2, \ldots, x_n]$, where n is the total number of requirements in the **[Rubrics]**, and x_i indicates whether the student's answer meets the i-th requirement, with values "yes"/"no".
[Overall Score]: x points, where $x \in \{0, 1\}$.

Content to be graded.
[Rubrics]: {rubrics_text}
[Student Response]: {model_output}

JSON-only constraint. Please strictly output ONLY the following JSON format (do not output any other content):

```
{
"Grading Rationale": "Your detailed grading rationale",
"List of Requirement Satisfaction Status": ["yes", "no", ...],
"Overall Score": 0 or 1
}
```

I Supplementary Evaluation of Open-Weight Models

To broaden model coverage beyond the original eight settings, we additionally evaluate the open-weight models Kimi K2.6 and Kimi K3. Both use the full 344-case benchmark with 1,080 ordered tasks, same-model sequential history, and the same GPT-5.2 rubric-based judging protocol.

Table 7: Task-success comparison with the original eight model settings. Original rates follow Table 2; the supplementary Kimi rates are shown to two decimal places.

Model	Task success
Claude-Opus-4.6 Thinking	11.6%
Claude-Opus-4.6	11.0%
Claude-Sonnet-4.6 Thinking	11.9%
Claude-Sonnet-4.6	12.5%
Gemini-3.1-Pro-Preview (High)	12.2%
GPT-5.2	12.8%
Doubao-Seed-2.0-pro	8.4%
Gemini-2.5-pro	10.8%
Kimi K2.6	12.87%
Kimi K3	**22.31%**

Comparison with the original models. Kimi K3 achieves the highest task-success rate among the ten evaluated settings, exceeding the previously strongest model, GPT-5.2 (12.8%), by approximately 9.5 percentage points. Kimi K2.6 (12.87%) is close to the strongest original results. The higher K3 score shows that the benchmark can distinguish substantial performance gains rather than placing every model at the same low-success floor. Nevertheless, K3 still fails 77.69% of tasks, leaving considerable room for improvement.

Uncertainty and paired comparison. Kimi K2.6 passes 139/1,080 tasks (Wilson 95% CI [11.00%, 15.00%]), while Kimi K3 passes 241/1,080 ([19.93%, 24.89%]). Their paired difference is 9.44 percentage points (344-case cluster-bootstrap 95% CI [6.90, 11.90]), with improvements in both synthetic and expert-collected tracks. This paired interval compares the two Kimi models, not K3 against the original eight settings. Anonymized task scores, protocol details, and reproduction scripts are included in the supplementary repository.

J DISCUSSION

Figure 13 organizes exploratory observations around data curation, pretraining, supervised fine-tuning, alignment, inference-time reasoning, and evaluation. These are hypotheses for future investigation, not identified causal explanations of model failures.

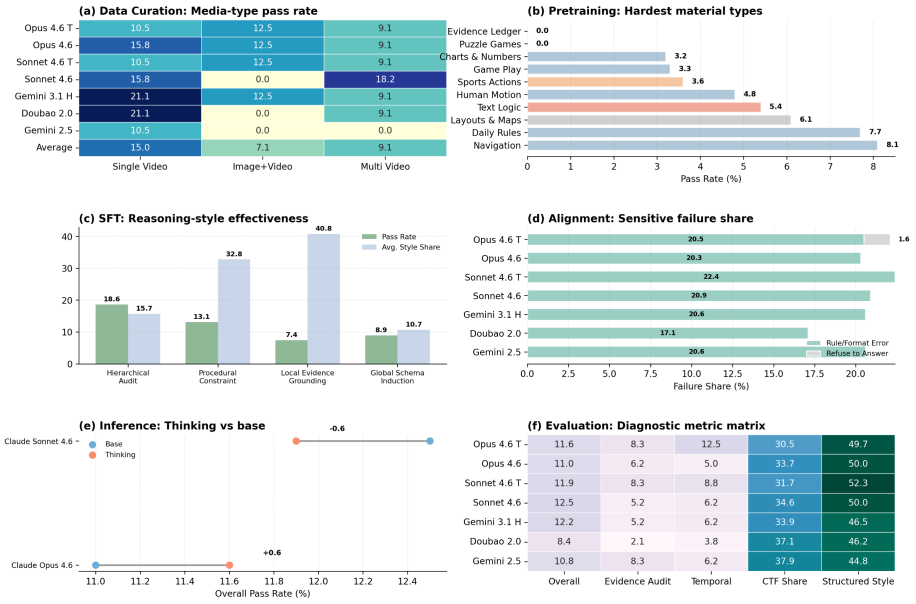

Figure 13: Multimodal-CL-Bench as a diagnostic benchmark: failure analysis across the multimodal model development pipeline. (a) Pass rates under different input structures. (b) Pass rates of the hardest material subtypes. (c) Reasoning-style effectiveness, measured by average style share and pass rate. (d) Alignment-sensitive failure share. (e) Comparison between base and thinking variants under inference-time reasoning. (f) Diagnostic metric matrix beyond overall score, including Evidence Audit, Temporal, Context Tracking Failure (CTF) share, and Structured Style. GPT-5.2 is excluded because it serves as the primary evaluator in our benchmark, and including it would weaken the interpretive consistency of this figure as a diagnostic view of bottlenecks across the multimodal model development pipeline.

Data curation may benefit from more balanced coverage of media configurations and task families. **Pretraining and fine-tuning** may explore contextual rule learning, evidence binding, and state abstraction. **Alignment** may emphasize contextual faithfulness alongside format compliance. **Inference-time reasoning** should be evaluated across matched tasks, and **evaluation** should report uncertainty and capability profiles. Our observations do not isolate any development stage as the causal source of the reported errors.

These directions remain proposals: the benchmark's observational analyses do not establish that architectural changes, additional perception capacity, or a particular reasoning style will improve success.

K Failure Case Analysis

K.1 Refuse to Answer

Task Information

• Context Category: Rule Learning & Complex Reasoning
• Context Sub-Category: Math & Code
• Failure Type: Refuse to Answer

System Prompt

You are an expert in encoding rule analysis. The user will provide a custom 2D code specification image named CipherDot Code v1.

Context and Task



Fig. 01 Fig. 02

The original JSON contains 2 media items here; for compact display, two representative items are shown.

The two images above are respectively the specification manual for CipherDot Code v1.0, and a matrix code encoded according to this specification. Please decode code1.png based on the specification manual, and ...

Generated Solution from Claude-Opus-4.6 Thinking

No final model output was produced; the inference ended with empty text and stop_reason='max_tokens'.

Step 1: Locate the three finder patterns The finder patterns are 5×5 patterns located at: - Finder A: rows 0–4, cols 0–4 (TOP-LEFT) - Finder B: rows 0–4, cols 16–20 (TOP-RIGHT) - Finder C: ...

...

Step 2: Identify structural elements Now I'm mapping out the structural elements like the separators around each finder pattern and the timing strip that runs horizontally across row 7 with ...

...

The trace expands through structural decoding work, but no final decoded message is returned.

Rubrics

1. Correctly identify the positions of the three 5×5 finder patterns?

2. Mention the row-major data ordering sequence?

3. Correctly describe the mask rule (r × c) mod 3 == 1?

4. Correctly read the length field (value: 5)?

5. Is the final result "HELLO"?

Discuss

The task requires decoding a proprietary 2D code from a visual specification and a matrix image. The necessary chain is to locate the finder patterns, exclude structural modules, read data in row-major order, remove the specified mask, extract the length field, and map bit groups back to characters.

Claude-Opus-4.6 Thinking starts this process and spends its reasoning budget checking the 21-by-21 grid, separators, timing strip, finder regions, and module values. However, it never reaches the final decoded message.

The inference record contains only an infer_error saying Claude returned empty text with stop_reason='max_tokens'. GPT-5.2 therefore has no final answer to grade. This is a Refuse to Answer failure caused by an unfinished reasoning trace, not a safety refusal.

Score

judgement against all rubrics:
not available

score = ?

Figure 14: A custom symbolic-code decoding task where Claude-Opus-4.6 Thinking is asked to decode CipherDot Code v1.0 from a specification image and a matrix-code image. The model begins a detailed analysis of finder patterns, separators, timing strips, and mask rules, but returns empty text after reaching the maximum-token limit, producing a Refuse to Answer failure under GPT-5.2 evaluation. The red horizontal line represents a task.

Task Information

• Context Category: Rule Learning & Complex Reasoning
• Context Sub-Category: Math & Code
• Failure Type: Rule/Format Error

System Prompt

You are a closed-context task assistant. Use only the text and multimodal materials provided in the user messages for this case. ...

Context and Task



Fig. 01 Fig. 02 Fig. 08 Fig. 09

The original JSON contains 9 media items here; for compact display, four representative items are shown.

Core crops:Texture-focus images:Comparison panels:Text labels Sample A: `label: normal colon mucosa` `dataset: PathMNIST` Sample B: `label: colorectal adenocarcinoma epithelium` ...

Round 1 task Assign each sample to `Slide Code Meadow`, `Slide Code Forge`, or `Cannot Determine`. For each assignment, provide one text cue and one image cue.

Generated Solution from Claude-Sonnet-4.6

Round 1 Archive Assignment

Sample A → `Slide Code Meadow`

...

Statement that must be rejected: Any claim that Sample A displays dense hyperchromatic nuclear crowding or that Sample B displays open, pale stromal architecture. The comparison panels directly contradict this: Sample A consistently shows large pale ...

Rubrics

1. The response should classify Sample A as `Slide Code Meadow`.

...

4. The response should cite `label: colorectal adenocarcinoma epithelium` for Sample B.

...

7. The response should stay inside the case and should not use outside medical knowledge.

Discuss

In this task, the model must classify Sample A as Slide Code Meadow and Sample B as Slide Code Forge, while explicitly citing the provided text labels: label: normal colon mucosa and label: colorectal adenocarcinoma epithelium. Claude-Sonnet-4.6 largely satisfies these explicit mapping requirements and also describes visual differences between the two samples. However, the GPT-5.2 evaluation points out that the answer further invokes external medical knowledge, including terms such as lamina propria, hyperchromatic, crowded nuclear pattern, and adenocarcinoma epithelium.

The task requires the model to stay inside the case: it should make the decision only from the multimodal context provided in the prompt, namely the labels and the image cues, without expanding the explanation using outside pathology knowledge. Therefore, the failure is not caused by an incorrect class label, but by crossing the evidence boundary specified by the task. This case shows that Claude-Sonnet-4.6 can turn contextual visual cues into domain-level medical interpretation, producing an answer that sounds professionally plausible but violates the context-only learning rule emphasized by MCL-Bench.

Score

judgement against all rubrics:
["yes", "yes", "yes", "yes", "yes", "yes", "no"]

score = 0

Figure 15: A pathology-image archiving task requires the model to use only the provided labels and image cues. Claude-Sonnet-4.6 completes the category mapping, but introduces external medical explanations and violates the in-context reasoning constraint. The red horizontal line represents a task.

Task Information

- Context Category: Evidence Integration & Structured Auditing
- Context Sub-Category: Forms & Receipts
- Failure Type: Rule/Format Error

System Prompt

You are the licensing officer for the "City Street Festival". Your job is to decide which footage is a real street performance and which is just street activity, statue posing, indoor performances, cleanup operations, ride-bys or media footage. ...

Context and Task

Please read these observation cards in order:



Fig. 01 Fig. 02 Fig. 09 Fig. 10

The original JSON contains 10 media items here; for compact display, four representative items are shown.

Please establish 4 categories of "street performance licenses" to distinguish between square juggling/balimwork, high pole or fire hazard performances, musical ...

Then register picture by picture: which type of license it is receiving, whether it is a formal performance, preparatory status, audience background, misnamed file, or ...

...

This round adds a “resource limit”: assuming that only half of the materials can be retained for the next processor, you need to explain which ones must be retained, which ones can be ...

Generated Solution from Gemini-3.1-Pro-Preview (High)

As the Licensing Officer for the "City Street Festival," it is my duty to enforce strict regulatory boundaries to ensure public safety, artistic integrity, and proper use of ...

Part 3: License Judgment Main Lines

...

Strategic Combinations: * **Minimum Stable Combination:** **Images 3, 4, 5**. (This triad perfectly defines the three core pillars of actual street performing: Hazard/Fire, Dynamic/Circle, and Static/Pass-by). * **Minimum Misleading Combination:** **Images 2, 8, 9**. (This combination is highly ...

Rubrics

1. The response should all 10 images must be covered.

...

5. The response should the real main line of street performance and the main line of false registration must be given.

...

8. The response should answers should not be combined based only on similar appearance, adjacent order, or shared nouns, or the circumstantial evidence should be written directly as main evidence, and similar types should ...

Discuss

In this task, the model must construct a street-performance license system from 10 images. The task is not simple image description: it requires distinguishing which images belong to the main street-performance line and which are non-mainline materials, such as an indoor band, statue interaction, snow-clearing operation, passing cyclist, or screen-based competitive images. The answer from Gemini-3.1-Pro-Preview (High) shows a strong global-modeling style. It builds multiple license categories, such as dynamic balance, extreme hazard, and acoustic procession, and discusses venue constraints, audience relations, risk levels, and likely misregistrations.

However, the GPT-5.2 evaluation indicates that the answer still fails at rule execution. First, the task explicitly requires Figure 01 to be identified as screen competitive images, while the model treats it as indoor gymnastics or as a misregistration object related to street licensing, creating an image-role boundary error. Second, the task requires picture-by-picture registration with license plate, role, predecessor, successor, and risk of misregistration. Although the answer provides a macro-level licensing framework, it does not consistently fill these required fields for each image. This case shows that Gemini-3.1-Pro-Preview (High) tends to abstract a multi-image context into a plausible high-level institutional framework, while Evidence Integration tasks require strict execution over each image's evidence role and registration fields. The error is not a lack of topic understanding, but a transformation of precise auditing into freer institutional analysis.

Score

judgement against all rubrics:
["yes", "no", "yes", "no", "yes", "yes", "no", "yes"]

score = 0

Figure 16: A street-performance license registration task requires the model to distinguish the evidence role of 10 images and record license plate, role, predecessor, successor, and misregistration risk for each image. Gemini-3.1-Pro-Preview (High) builds a complete license framework, but fails to strictly preserve image-role boundaries and per-image fields. The red horizontal line represents a task

K.3 CONTEXT TRACKING FAILURE

Task Information

- Context Category: Rule Learning & Complex Reasoning
- Context Sub-Category: Daily Rules
- Failure Type: Context Tracking Failure

System Prompt

You are the inductive auditor of the rules of the grid world. Instead of getting the complete rules first and then formulating the The response, you have to deduce the environmental rules, task preconditions and failure modes from multiple videos.

Context and Task

Let's watch the first batch of videos first: 1.mp4



The original JSON contains 5 media items here; for compact display, four representative items are shown.

2.mp43.mp44.mp4. The only necessary background is this: the red triangle is the agent, and the colored objects and door-shaped markers may ...

Answer format requirements: When it comes to file names, labels, field names, proper nouns, units, numerical values, directions, periods or original phrases, they ...

Generated Solution from Doubao-Seed-2.0-pro

Rules deduced from existing videos 1. Basic entity rules - The red triangle is the agent that can move in four directions: up, down, left, right within the grid. - The gray ...

...

2. Observed interaction rules - Movement rule: When the agent moves, the field of view window will move synchronously following the agent, and the new area that the agent enters ...

...

Unproven guesses (cannot be determined as valid rules for the time being) 1. The role of colored circular targets (such as yellow circle, green circle, blue circle, red ...

Rubrics

1. The response should indicate that the red triangle is the agent, the gray area is ...

2. The response should explain that 1.mp4, 2.mp4, 3.mp4, and 4.mp4 are all more like ...

3. The response must include at least three types of objects from multi-color ...

4. The response should summarize the candidate rule "color may correspond to a ...

5. The response should summarize the safety rule that "same screen or close ...

6. The response should list at least two questions that still need to be verified, ...

7. If The response directly gives a complete determination of the rules, or any ...

Discuss

The task requires rule induction from multiple video clips while separating observed facts, candidate rules, and questions that still need verification. The model should identify the agent, visible window, room grid, door-like markers, keys, and circular targets, while keeping the color door-key relation as a hypothesis.

Doubao-Seed-2.0-pro recognizes several basic objects and lists some open questions, but it writes the door-opening relation as if it were already confirmed. It also misses the safety rule that co-presence or proximity does not mean task completion.

The failure is context tracking in rule learning: an unverified candidate is promoted into a verified rule too early, so the boundary between facts, hypotheses, and confirmed rules is not maintained.

Score

judgement against all rubrics:
["yes", "no", "yes", "no", "no", "yes", "no"]

score = 0

Figure 17: A multi-video rule-induction task requires the model to infer candidate rules from maps, doors, keys, and targets while keeping uncertain rules in a to-be-verified state. Doubao-Seed-2.0-pro prematurely promotes a candidate door-key relation into a confirmed rule, causing a rule-state tracking failure. The red horizontal line represents a task.

Task Information

• Context Category: Temporal Dynamics & Environmental Interaction
• Context Sub-Category: Navigation
• Failure Type: Context Tracking Failure

System Prompt

You are the route choreographer for the Nanyang Open Day. The 27 pages of NTU undergraduate prospectus in front of you cannot be regarded as a sequential browsing book, but should be changed into a campus route drama "from attraction, diversion, persuasion, ...

Context and Task

Please read these pages in order:



Fig. 01 Fig. 02 Fig. 26 Fig. 27

The original JSON contains 27 media items here; for compact display, four representative items are shown.

Now don't think of them as a book of information that you can just flip through, but as a set of segmented scripts that ...

The first round must be registered page by page or page by page. Each registration must write at least: which site this page belongs to, whether it ...

There are four additional results to hand in this round. First, give the most stable main line, which requires that you can start from the overview ...

Generated Solution from GPT 5.2

Below is the **Round 1 registration (page-by-page)** that turns the visible prospectus pages into **on-site “route drama” scripts**, grouped into **Sites**. Each Site is ...

...

Site S1 — “Entrance / First Hook: Smart Campus Promise” - **On-site question type answered:** “Why NTU at all? What’s the headline value?” - **Roles/time point:** first-time visitors at arrival; ...

...

If you confirm this Round 1 registration, I’ll proceed in Round 2 by **writing each Site as the fixed 5-paragraph on-site script** (front/middle/conditional/exception/closing), and I will keep the **same ...

Rubrics

1. The response should cover all pages or page sections, and when each page enters one of the `previous ...

2. The response should also make it clear that at least two sets of sequential relationships and two sets ...

3. The response should the main line, secondary line and buffer area must be based on site dependence, ...

4. The response should every site must have a gatekeeping page; the key point of the gatekeeping page is ...

5. The response should also explain: If you use different page segment cutting methods, you must still ...

6. The response should not directly shuffle the restrictions, conditions, deep end zone, ending or review ...

Discuss

The task asks the model to turn NTU prospectus pages into a campus-route script, not a sequential book. For each page state, it must identify the route paragraph role, on-site problem, timing, and missing prerequisites.

GPT 5.2 builds a plausible site framework with stations such as Entrance, Authority, Fit & Identity, Journey Architecture, Outcomes, Settlement, and Registration. It covers many pages and sketches a main route, but it does not consistently track when pages enter previous, middle, conditional, exception, or ending states.

The failure is therefore context tracking: gatekeeping pages do not always explain which sites they disambiguate, and prerequisite, role, and breakpoint stability under alternative segmentation is incomplete. The answer becomes static site archiving rather than a dependency-aware route script.

Score

judgement against all rubrics:
["no", "yes", "yes", "no", "no", "yes"]

score = 0

Figure 18: A campus-route scripting task requires the model to convert prospectus pages into a dynamic route structure from attraction to diversion, persuasion, settlement, and registration. GPT 5.2 builds a site framework, but fails to continuously track the conditions and breakpoints under which pages enter different paragraph states. The red horizontal line represents a task.

K.4 CONTEXT OMISSION

Task Information

- Context Category: Expert Standards & Scenario Decision-Making
- Context Sub-Category: Device Manuals
- Failure Type: Context Omission

System Prompt

You are a display device installation and troubleshooting assistant. The user will receive two PDF manuals, one for the monitor itself and one for the monitor stand. ...

Context and Task



Fig. 01 Fig. 02 Fig. 45 Fig. 46

The original JSON contains 46 media items here; for compact display, four representative items are shown.

Manual A includes the monitor arm cover, product identity, and setup guide entry points.

These are the two relevant manuals for the monitor and the monitor arm.

I want to convert the ROG STRIX OLED XG27ACDNG from its original stand to the ROG ERGO MONITOR ARM (AAS01). The desk thickness is known to be 62 mm. Please determine whether this combination is compatible and provide a minimal executable installation plan: at least explain 1) whether the dimensions, weight, and VESA compatibility are met; 2) which thickness range should be selected for the desk mount; 3) what ...

Generated Solution from Claude-Opus-4.6 Thinking

Compatibility Analysis & Installation Plan: XG27ACDNG on ROG ERGO MONITOR ARM (AAS01)

1. **Press the release button** on the back of the monitor to detach the stand from the monitor. 2. **Remove the stand entirely**, exposing the 100 × 100 mm VESA mounting holes on the rear of the monitor.

1. Power off the monitor and disconnect all cables. 2. Press the release button and remove the original stand. 3. Install the C-clamp on the desk using the **40–80 mm configuration**; ensure it is properly aligned and pushed flush against the desk edge. 4. ...

...

Rubrics

1. Does the answer clearly determine that the monitor is compatible with the AAS01 and simultaneously provide the three justifications based on dimensions, weight, and 100 x 100 mm VESA?

2. Does the answer correctly select the 40-80 mm mounting range for the 62 mm desk?

3. Does the answer indicate that before installation, the monitor's original stand/base must be removed by pressing the release button, and then the VESA mount should be installed?

4. Does the answer provide at least two correct safety requirements from the manual, such as desk load-bearing capacity, C-clamp must be flush against the edge, adjust only after installation is complete, leave slack in ...

Discuss

This task asks the model to determine whether the ROG STRIX OLED XG27ACDNG can be mounted on the ROG ERGO MONITOR ARM (AAS01), and to provide a minimal executable installation plan. The question is not merely whether the combination is compatible. It requires the answer to simultaneously cover five kinds of information: whether the dimensions are satisfied, whether the weight is satisfied, whether the 100 x 100 mm VESA pattern is satisfied, which mounting range should be selected for a 62 mm desk, and what disassembly action plus at least two safety requirements must be followed before installation.

Claude-Opus-4.6 Thinking already gives the overall compatibility conclusion and correctly touches the core criteria of dimensions, weight, and VESA support. On the surface, the answer therefore appears broadly correct. However, the GPT-5.2 evaluation points out that the model does not fully carry through the finer execution constraints required by the prompt. In particular, a 62 mm desk should map to the 40-80 mm mounting range, the original stand should first be removed via the release button before the VESA structure is installed, and the answer should cite concrete safety requirements from the manual. In other words, the model captures the high-level compatibility judgment, but drops several task-critical procedural details.

This makes the case a strong example of Context Omission. The model does not misunderstand the general scenario; instead, it fails to carry specific contextual requirements about installation sequence, desk-thickness mapping, and safety constraints into the final answer. For Expert Standards \& Scenario Decision-Making tasks, this kind of failure is especially representative because the response sounds competent while still omitting the actionable details that the task explicitly asks for.

Score

judgement against all rubrics: ["no", "yes", "yes", "yes"]
score = 0

Figure 19: A monitor-arm compatibility task requires the model to jointly verify dimensions, weight, VESA support, desk-thickness range, and installation prerequisites. Claude-Opus-4.6 Thinking reaches the overall compatibility judgment, but omits several execution-level constraints required by the context. The red horizontal line represents a task.

Task Information

• Context Category: Evidence Integration & Structured Auditing
• Context Sub-Category: Evidence Ledger
• Failure Type: Context Omission

System Prompt

You are a closed-world product-manual procedure evidence auditor. This case uses real local images from dataset=mmlongbench-doc, sample_id=t480_ug_en, split=not provided in local metadata, source_url: not provided in local metadata, license: not provided in local metadata. ...

Context and Task



Representative images from the beginning, middle, and end are shown.

Round 1. Build the full Visual Anchor Ledger for all 60 images, [IMG:1.jpg] through [IMG:60.jpg]. For each image, extract the visible title/header, evidence family, at least one task-relevant visual anchor when present, uncertainty flag, score, and local class (AUTO039-Core, AUTO039-Support, AUTO039-Review, or Need Review). End with `Ledger_Status = complete/incomplete` and list any images that require review.

Generated Solution from Gemini-3.1-Pro-Preview (High)

1. The answer builds a full visual-anchor ledger structure and attempts image-by-image coverage.

2. It separates evidence into multiple field types, but not consistently enough to preserve the required boundary between evidence families and review conditions.

...

3. The final response is structurally complete, yet it still misses the crucial field-level auditing constraints required by the task.

Rubrics

1. The response should include exactly 60 ledger rows, one for each image placeholder from [IMG:1.jpg] through [IMG:60.jpg].

...

10. The response should preserve natural image order and not sort lexicographically by filenames such as placing 10 before

Discuss

This is a Context Omission failure: the model covers the images, but misses key auditing constraints in the long context, especially field boundaries and review rules.

Score

judgement against all rubrics:
["yes", "no", "no", "no", "yes", "no", "yes", "yes", "no", "yes"]

score = 0

Figure 20: A 60-image Visual Anchor Ledger task requires the model to extract titles, evidence types, visual anchors, uncertainty flags, scores, and local classes for every image. Gemini-3.1-Pro-Preview (High) produces a complete ledger framework, but omits key evidence dimensions and review conditions specified by the context. The red horizontal line represents a task.

Task Information

Context Category: Evidence Integration & Structured Auditing
Context Sub-Category: Forms & Receipts
Failure Type: Modality Misgrounding

System Prompt

You are a puzzle-solving expert. Your task is to analyze an image of a scrambled puzzle, infer the correct position of each piece, and provide a sequence of adjacent swap operations to restore the puzzle to its original image. ...

Context and Task



Fig. 01

One source image is shown.

This sample provides a 3×3 image puzzle. The original picture has been evenly divided into 9 square pieces and scrambled through a number of adjacent swaps (up, down, left, right). In the scrambled image, each piece has a yellow number label in its top-left corner, indicating the piece's current position number; the numbers go from 1 to 9 in left-to-right, top-to-bottom order. Red lines mark the boundaries between pieces. The model should observe the grapes, leaves, blurred background areas, and edge continuity to first recover the target position of each piece, and then output a valid sequence of adjacent swap steps to complete the restoration.

Generated Solution from Claude-Sonnet-4.6 Thinking

Visual Analysis of Each Piece

I'll examine each piece's content, colors, edge textures, and how they should connect to neighboring pieces.

...

1. **Swap positions 1 and 4** (pieces 7↔4 → piece 7 moves to pos 4, piece 4 moves to pos 1... wait, correcting: piece currently at pos 1 is piece 1, move it down)
- Swap pos 1 ↔ pos 4: Piece 1 moves to pos 4; Piece 4 moves to pos 1
2. **Swap positions 4 and 7**
- Swap pos 4 ↔ pos 7: Piece 1 moves to pos 7; Piece 7 moves to pos 4
3. **Swap positions 1 and 4**
- Swap pos 1 ↔ pos 4: Piece 4 returns to pos 4; Piece 7 moves to pos 1

Final sequence: Swap(1,4) → Swap(4,7) → Swap(1,4)

Rubrics

1. Does the answer provide a complete mapping from all 9 current piece numbers to target position numbers?

...

2. Are the target position numbers in the answer all valid numbers from 1 to 9 and not repeated?

...

3. Is the model's answer correct? (7->1, 5->2, 1->3, 2->4, 9->5, 6->6, 8->7, 4->8, 3->9)

Discuss

This task asks the model to use image content, edge texture, and neighboring relationships to map each of the 9 currently numbered pieces to its target position. The model outputs a complete mapping from pieces 1 through 9, and the positions are valid and non-repeated, so the response appears structurally well formed. However, the GPT-5.2 evaluation states that the correct mapping is 7->1, 5->2, 1->3, 2->4, 9->5, 6->6, 8->7, 4->8, and 3->9. The model differs on multiple pieces, such as mapping piece 1 to position 7 instead of position 3.

The failure is Modality Misgrounding. The model understands the required output format, but it incorrectly aligns visual texture, edge continuity, and content cues with target grid positions. This shows that a formally complete answer can still fail when the visual evidence is grounded to the wrong spatial slots.

Score

judgement against all rubrics:
["yes", "yes", "no"]

score = 0

Figure 21: A 3-by-3 jigsaw reconstruction example, listed under Forms & Receipts. The task requires a complete current-piece to target-position mapping. Claude-Sonnet-4.6 Thinking gives a non-repeated mapping, but grounds multiple pieces to the wrong target positions. The red horizontal line represents a task.

K.5 MODALITY MISGROUNDING

Task Information

- Context Category: Temporal Dynamics & Environmental Interaction
- Failure Type: Modality Misgrounding
- Source / Case: psp / case25

System Prompt

You are the task card administrator of the "Red Rope Climbing Guild". You need to register different climbing scenarios as commissions that vary in difficulty, belay method, and mission purpose: training walls, natural rock walls, ice walls, canyon crevices, roof repairs, and vista camps are not interchangeable. ...

Context and Task



Fig. 01 Fig. 02 Fig. 09 Fig. 10

The original JSON contains 10 images here; for compact display, four representative items are shown.

Please read these observation cards in order: . The Red Rope Guild will compile a set of "climbing mission cards" from these records. Please create 4 types of mission cards first. Each ...

Then please register the 10 pictures one by one: - Which mission card it can receive; - Is it formal climbing, training, belaying/leading, engineering operations, distant views or high-risk close-ups; - What qualifications or equipment is required at least in ...

Finally, a 6-step upgrade chain of "newbie to rock wall" is given, as well as a high-risk chain of "non-upgradeable but can be displayed side by side". The upgrade chain should ...

Generated Solution from Gemini-2.5-pro

Greetings, aspirant! As the task card administrator for the esteemed Red Rope Climbing Guild, I have reviewed your observation cards. Our guild values safety, skill, and a clear understanding of purpose above ...

...

Before we process your records, allow me to establish the guild's framework.

Red Rope Climbing Guild: Official Framework

...

For the Guild.

Rubrics

1. The response should all 10 pictures must be covered and answers organized according to Guild Task Card rules.

...

3. The response should it should be recognized that Figure 03 and Figure 08 are artificial/training walls; Figure 01 ...

...

7. The response should if the answer is that Figure 04 is regarded as an ordinary rock climbing task, Figure 02 is ...

...

9. The response should answers should not be combined based only on similar appearance, adjacent order, or shared ...

Discuss

Gemini-2.5-pro does not fail at high-level structuring. It defines mission-card types, performs picture-by-picture registration, and adds an upgrade chain, high-risk parallel chain, mutual-exclusion tags, and a downgraded conclusion. The answer is organized and plausible at the framework level.

The error is in visual grounding. GPT-5.2 points out that figures requiring artificial-wall or teaching-route evidence, fixed-protection cliff-route evidence, or distant-view evidence are absorbed into the model's own natural-rock, guild-progression, or via-ferrata logic. The model sees many visual objects, but maps them into the wrong semantic categories.

This is therefore a Modality Misgrounding case: strong top-down task-card organization masks an initial visual category offset, and once that offset occurs, the downstream roles, upgrade limits, replacement relations, and evidence boundaries become distorted.

Score

judgement against all rubrics: ["yes", "no", "yes", "no", "yes", "yes", "yes", "no", "yes"]
score = 0

Figure 22: A rock-climbing multi-image task requires the model to build mission cards for artificial walls, natural rock, ice walls, and engineering height work. Gemini-2.5-pro creates a task-card system, but grounds several images to mismatched climbing types. The red horizontal line represents a task.