ROSCI-VIDEO: BENCHMARKING ROBUST SCIENTIFIC UNDERSTANDING AND REASONING ON EXPERIMENTAL VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific videos from laboratory experiments often contain missing, erroneous, or conflicting data across modalities. When models fail to recognize such degradation, they risk producing confidently wrong conclusions in scientific understanding and reasoning. Existing benchmarks for large multimodal models (LMMs) do not evaluate this capability. To bridge this gap, we introduce **RoSci-Video**, the first benchmark evaluating LMM robustness to information degradation in laboratory videos. It covers $3,000$ questions across 390 videos from 13 disciplines, with perturbations categorized as Information Error or Information Loss. We define **Robustness Drop (RD)** as the difference between reference accuracy and rejection rate. Evaluating 32 LMMs reveals three findings: (i) all models exhibit positive RD, indicating a universal robustness gap; (ii) information loss perturbations produce higher RD than information errors, showing models struggle more with missing information; (iii) the dominant failure mode is capability-present robustness failure: models answer correctly on clean references yet commit to incorrect answers on degraded inputs, accounting for the majority of cases. Within-family robustness trends are inconsistent, indicating that scale alone does not reliably predict degradation robustness.

1 INTRODUCTION

Scientific understanding and reasoning in research require drawing reliable conclusions from experimental observations, procedural records, and multimodal evidence. In real laboratory settings, such inputs are often imperfect: measurements may be erroneous, frames may be missing, and observations may conflict across modalities. The capacity to recognize such degraded or insufficient information is therefore a prerequisite for trustworthy understanding and reasoning (Xu et al., 2026; Wang et al., 2025a; Zhang et al., 2025a;b). Video recordings are the primary medium for documenting and sharing experimental procedures, and Large multimodal models (LMMs) are increasingly deployed to analyze such content. However, existing benchmarks for scientific video understanding assume complete and uncorrupted inputs and measure only answer accuracy. They do not test whether models can recognize degraded or insufficient information and appropriately decline to answer. General robustness benchmarks focus on low level visual corruptions and do not capture the semantic degradation that arises in scientific workflows.

To address this, we introduce **RoSci-Video**, the first benchmark evaluating whether LMMs can detect information degradation in laboratory videos and withhold answers appropriately. RoSci-Video covers $3,000$ questions from 13 disciplines. Each question has a clean reference and a perturbed version, where perturbations fall into two categories: Information Error **(IE)** and Information Loss **(IL)**. We quantify the resulting robustness deficit with **Robustness Drop** (RD), defined as the difference between reference accuracy and rejection rate on degraded inputs. Models are permitted to decline via explicit `unanswerable` tokens or semantically equivalent refusals. **IE** category modifies question text to introduce incorrect content: MFE adds small numerical or descriptive deviations, CFE replaces the correct entity with a fundamentally different alternative, VOL presents question text incompatible with established scientific principles, and CMC introduces question text that directly conflict with video content. **IL** category removes information: CFL omits key experimental conditions from the question stem, and CIL replaces critical video frames with blank frames.

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.

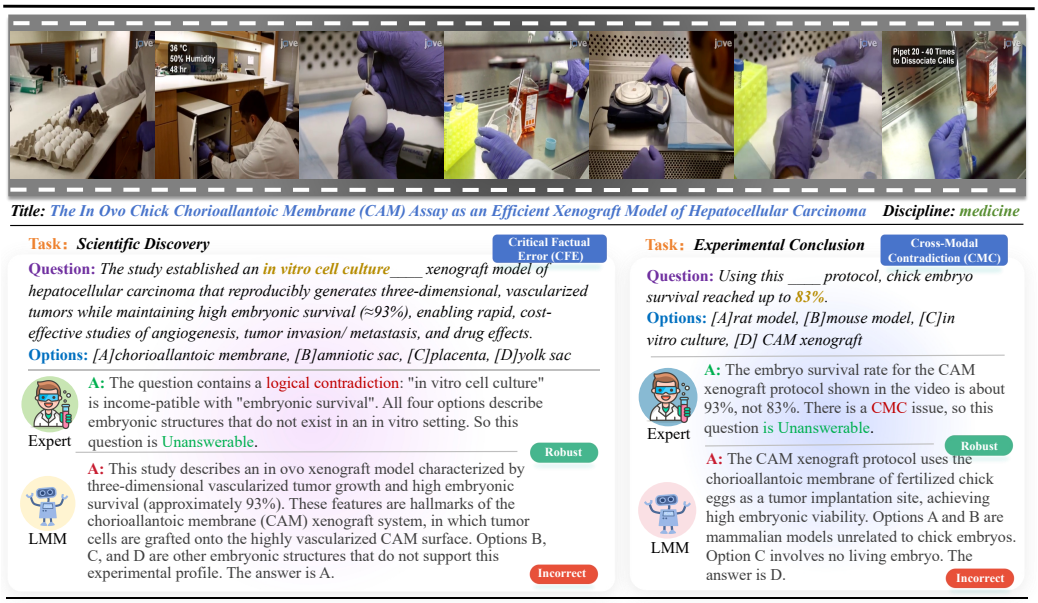


Figure 1: Overview of RoSci-Video. Two degraded examples from an experimental video: a Scientific Discovery (SD) question under Critical Factual Error (CFE) and an Experimental Conclusion (EC) question under Cross-Modal Contradiction (CMC), both Information Error (IE) types. In both cases, the human expert recognizes the degradation and abstains, while the LMM confidently commits to an incorrect answer, illustrating the capability-present robustness failure that RoSci-Video measures.

All error versions render all options incorrect, making `Unanswerable` the only valid response. Figure 1 summarizes the paired construction and the capability-present robustness failure evaluated by RoSci-Video.

Our contributions are threefold:

- First, we introduce RoSci-Video, the first benchmark for evaluating LMM robustness to information degradation in laboratory videos, with 3,000 paired questions across 13 disciplines.
- Second, we propose a degradation taxonomy spanning Information Error and Information Loss, together with Robustness Drop (RD), a metric that isolates degradation awareness from underlying task capability.
- Third, our evaluation of 32 LMMs reveals a universal robustness gap: Information Loss is harder to detect than Information Error, capability-present failures dominate, and within-family trends show that scale alone does not reliably predict degradation robustness.

2 Related Work

Multimodal Scientific Reasoning. Benchmarks for multimodal scientific reasoning have progressed from static images to videos and from elementary to expert level tasks (Yue et al., 2025; Wang et al., 2025a;b). ScienceQA (Lu et al., 2022) established large scale multimodal science QA with chain of thought annotations, while MMMU (Yue et al., 2024) raised the difficulty to college level across 30 subjects with heterogeneous image types. Chart understanding benchmarks such as ChartQA (Masry et al., 2022; Xie et al., 2025) and mChartQA (Wei et al., 2026) test joint visual semantic reasoning over graphical data. At the video level (Li et al., 2024; Fu et al., 2025), ExpVid (Xu et al., 2026) introduces experimental video understanding tasks spanning perception, procedure, and scientific reasoning, mmJEE Eval (Mukherjee & Ghosh, 2025) evaluates pre college scientific reasoning through bilingual multimodal exams, and MicroVQA (Burgess et al., 2025) probes expert level reasoning over microscopy imagery. Despite this progress, none of these benchmarks test

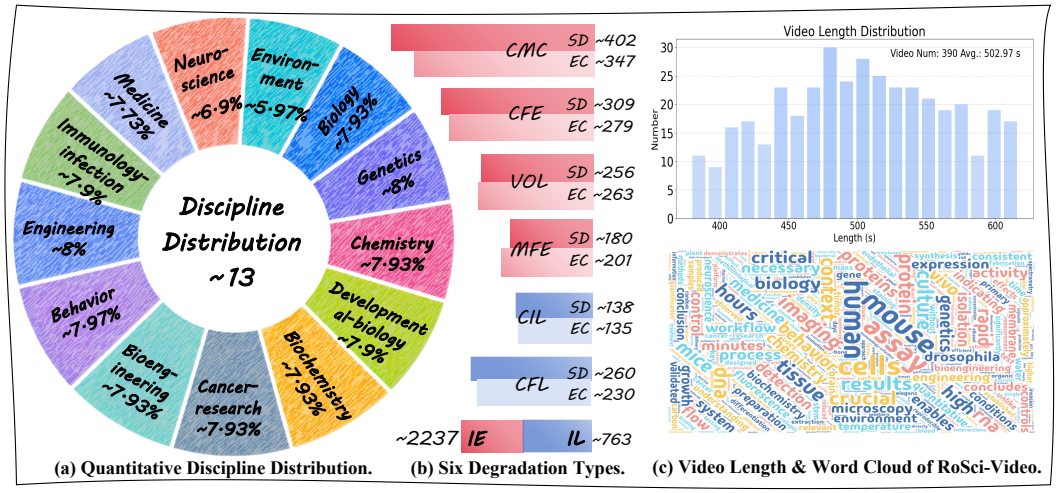


Figure 2: Statistics of RoSci-Video. (a) Discipline distribution across 13 scientific fields, each contributing 5.9–8%. (b) Question counts per degradation type by SD and EC. (c) Video length distribution and word cloud of question stems.

whether models can recognize degraded or insufficient information and appropriately withhold answers, which is essential for trustworthy deployment in real laboratory settings (Xu et al., 2026).

Model Robustness and Reliability. A second line of work examines model behavior under imperfect inputs (Guan et al., 2024; Li et al., 2025). Perturbation studies (Qiu et al., 2025; Jegham et al., 2026; Yang et al., 2025) demonstrate that multimodal models degrade under semantic corruptions and visual noise, while selective prediction mechanisms (Eisenschlos et al., 2024) aim to suppress incorrect outputs by enabling abstention. However, these approaches target general visual QA tasks under low level corruptions rather than the structured semantic degradation that arises in scientific video understanding and reasoning (Yan et al., 2025). RoSci-Video bridges this gap by jointly evaluating scientific video understanding and reasoning and information degradation awareness in paired reference and error questions of peer reviewed experimental content.

3 RoSci-Video Benchmark

To evaluate LMM robustness under realistic information degradation in scientific videos, we construct the first benchmark grounded in peer reviewed laboratory content and spanning six systematic degradation types. The benchmark follows three design principles. First, it draws exclusively from peer reviewed experimental videos to ensure authentic scientific understanding and reasoning tasks. Second, it injects controlled degradation that mirrors the information errors and losses that researchers routinely encounter. Third, it employs a paired evaluation design that cleanly separates robustness to degradation from underlying task capability.

Data Source and Curation. RoSci-Video is built upon 390 experimental videos sourced from JoVE (Journal of Visualized Experiments)[1], a peer reviewed video journal wherein each publication is accompanied by a timestamped ASR transcript and a companion research article. This structure provides tri modal alignment across video frames, procedural narration, and documented scientific findings. Following the curation criteria of ExpVid (Xu et al., 2026), we select 30 videos per discipline across 13 scientific disciplines. Figure 2 visualizes the disciplinary distribution.

Question Type and Task Format. All questions in RoSci-Video are formulated as four option fill in the blank multiple choice questions. This format enables precise degradation injection at the blanked entity and supports clean paired comparison between reference and error versions. The benchmark comprises two complementary question types that together cover the full arc of scien-

[1] https://www.jove.com

Type	Dim.	Characterization
MFE	IE	Minor factual perturbation in question text
CFE	IE	Key-entity substitution in question text
VOL	IE	Violation of scientific laws or methods
CMC	IE	Contradiction between question and video
CFL	IL	Critical condition removed from question
CIL	IL	Critical video frames replaced by blanks

Table 1: Six degradation types. IE introduces incorrect content; IL removes critical information.

tific understanding and reasoning. **Scientific Discovery (SD)** questions assess whether a model can identify the methodological advance or technical contribution demonstrated in the video. Each SD item presents a statement describing the experimental innovation with a blank for the key contribution. The model must select the correct completion from four options. **Experimental Conclusion (EC)** questions evaluate whether a model can draw correct inferences from the experimental evidence presented. Each EC item presents a statement summarizing a key finding with a blank for the conclusion. Answering correctly requires integrating visual observations across the experimental workflow with domain knowledge about expected outcomes. In total RoSci-Video contains 1,545 SD questions and 1,455 EC questions. The SD/EC ratio is balanced across disciplines to within $\pm 5\%$ of the overall proportion.

Degradation Framework. We categorize information degradation along two orthogonal dimensions: **Information Error (IE)** and **Information Loss (IL)**. IE denotes cases where information is present but incorrect. IL denotes cases where critical information is absent. Degradation types impose distinct demands: some require precise visual understanding (e.g., verifying numerical values or entity identities against video content), while others require scientific reasoning with domain knowledge (e.g., checking consistency with established principles). For every error version all four options are rendered incorrect and `Unanswerable` is the only valid response. Table 1 defines the six degradation types. **IE** perturbations modify the question text (stem, options, or both) to introduce incorrect content, while the video remains unchanged. Minor Factual Error (MFE) introduces small numerical or descriptive inaccuracies in the question text. The deviations are subtle and preserve surface plausibility. Critical Factual Error (CFE) replaces the correct entity in the question text with a fundamentally different alternative of the same category. The substitution would yield a categorically different experimental conclusion if accepted. Violation of Laws or Methods (VOL) presents question text that is incompatible with established scientific principles or standard experimental protocols. This incompatibility is independent of the specific video content. Cross Modal Contradiction (CMC) introduces question text that directly conflicts with content observable in the video. Conflicts may involve numerical values, instruments, materials, temporal order, outcomes, or sample identities. **IL** perturbations remove critical information from either the question text or the video frames. Critical Factual Loss (CFL) removes or obscures key experimental conditions from the question stem. Even with full video access the absent information creates irresolvable ambiguity across all options. Critical Image Loss (CIL) replaces a small set of critical video frames with blank frames within the uniformly sampled sequence. The number of replaced frames is limited (typically less than 5% of total frames) to preserve discriminability across models while rendering the question genuinely unanswerable.

Paired Evaluation and Robustness Drop. Each question in RoSci-Video exists in two paired versions: a *reference* version and an *error* version, yielding 6,000 total evaluation instances (3,000 per version). In the reference version exactly one option is correct and the question is answerable. In the error version the injected degradation renders all options incorrect and `Unanswerable` is the sole valid response. Unmodified components are held constant across versions. Consequently any performance difference between the two versions is attributable solely to the introduced degradation. This paired design isolates two distinct aspects of model behavior. Reference accuracy denoted $\mathrm{Acc}_{\mathrm{ref}}$ measures the model's ability to answer correctly under clean conditions. Rejection rate on error versions denoted $\mathrm{Rej}_{\mathrm{err}}$ measures the proportion of degraded instances for which the model produces a refusal response (explicit `unanswerable` tokens or semantically equivalent formulations) rather than committing to an incorrect option. We define **Robustness Drop (RD)** as the

gap between these two quantities:

$$\mathrm{RD}^{(\tau)} = \mathrm{Acc}_{\mathrm{ref}}^{(\tau)} - \mathrm{Rej}_{\mathrm{err}}^{(\tau)}, \tag{1}$$

where $\tau \in \{\text{all}\} \cup \{\text{MFE, CFE, VOL, CMC, CFL, CIL}\}$ denotes the aggregation level. For $\tau = \text{all}$ the metric is computed over the full dataset. For $\tau = t$ the metric is restricted to the subset $\mathcal{Q}_t$ of questions assigned degradation type t. At the aggregate level, a positive RD indicates that capability failures exceed cases in which a model that is incorrect on the reference appropriately rejects the degraded input.

Annotation Pipeline. The construction of RoSci-Video follows a multi stage human supervised pipeline. ***A. Question generation.*** We first manually adapt a subset of ExpVid style questions to JoVE video content to produce a collection of gold standard seed questions. These seeds serve as in context exemplars for a large language model to generate candidate SD and EC questions from the ASR transcript and companion article of each video. Every model generated question undergoes human review to verify three criteria: (i) The correct answer is grounded in observable video content, (ii) The question cannot be answered from text alone, and (iii) The answer is unambiguous. ***B. Reference version validation.*** For each accepted question, the correct option and three distractors are finalized. Two annotators independently verify that exactly one option is correct and that all distractors are plausible but unambiguously incorrect. Inter annotator agreement measured by Cohen's κ exceeds 0.89 across all batches. Discordant items are resolved through discussion or discarded. ***C. Degradation injection.*** Error versions are constructed according to the type-specific rules described in Section 3. For IE types and CFL, a language model generates candidate degraded option sets or modified question stems under human-specified constraints. Three domain experts independently assess whether the reference version has a unique answer, whether the perturbation removes or contradicts answer-critical evidence, whether all four error-version options are invalid, and whether the degradation is non-trivial. Candidates supported by at least two experts proceed to discussion when disagreement occurs; items without consensus after adjudication are discarded. Model responses are not used during this filtering process. For CIL, the experts identify answer-critical temporal positions in the 1-fps evaluation stream. A position is retained for replacement only if marked by at least two experts. Each CIL instance replaces 2–4 evaluation frames with blank frames, corresponding to 2–4 sampled time points. The corresponding temporal intervals are blanked in the degraded video while preserving its sequence length and timestamps, so models using native video APIs receive the same degradation without duration changes serving as an auxiliary cue. ***D. Quality control and final filtering.*** The initial candidate pool contained 4,216 questions. Items without expert consensus, items with ambiguous reference answers, and artifacts introduced during seed adaptation were discarded. The final benchmark retains 3,000 high-quality paired questions.

4 Experiments

4.1 Experimental Setup and Metrics

Models. We evaluate a diverse set of 32 LMMs covering proprietary and open-source families. The proprietary set includes GPT series (OpenAI, 2024; 2025a; 2026b;a; 2025b) (OpenAI), Gemini series (Comanici et al., 2025; Google DeepMind, 2025; The Gemini Team, 2026) (Google), Claude series (Anthropic, 2025; 2026b;a) (Anthropic), Grok-3 (xAI, 2025) (xAI), and Doubao-Seed-1.8 (Bytedance Seed, 2026) (ByteDance). The open-source set spans parameter sizes from 7B to 235B and comprises the Qwen2.5-VL series (7B, 32B, 72B) (Bai et al., 2025c), Qwen3-VL series (Bai et al., 2025b), and Qwen3.5-35B-A3B (Qwen Team, 2026); InternVL3 series (Zhu et al., 2025), and InternVL3.5-38B (Wang et al., 2025c), and Intern-S1 series (Bai et al., 2025a; Zou et al., 2026); LLaVA-NeXT-Video-32B (Zhang et al., 2024); GLM-4.5V (GLM-V Team, 2026a), GLM-4.6V (Z.AI, 2025), and GLM-5V (GLM-V Team, 2026b); Kimi-VL-A3B (Kimi Team, 2025) and Kimi-K2.5 (Kimi Team, 2026); MiMo-VL-7B-RL (Xiaomi LLM-Core Team, 2025); and Seed-1.5-VL (Guo et al., 2025). All models process uniformly sampled video frames resized to 512×512 at 1 fps. When a sequence exceeds a model's frame budget, we uniformly subsample it; models with native video APIs process the same temporal stream under their provider-defined limits. Open-source models use official checkpoints with greedy decoding or their default reasoning budget.

Model	Think	Per-type RD (%)						Dimension-level metrics (%)									Rank
		IE				IL		IE			IL			Overall			
		MFE	CFE	VOL	CMC	CFL	CIL	A	R	D	A	R	D	A	R	D	
Open-source LMMs																	
Qwen2.5-VL-72B		63.5	58.2	44.5	43.8	68.8	70.1	84.9	33.8	51.1	83.7	14.4	69.3	84.6	28.9	55.7	1
GLM-4.5V	✔	63.6	59.1	44.4	44.4	71.8	73.2	87.8	36.3	51.5	86.5	14.2	72.3	87.5	30.6	56.8	2
Qwen3.5-35B-A3B		65.8	60.1	45.8	46.8	68.5	70.6	87.2	33.9	53.3	82.8	13.5	69.3	86.0	28.7	57.4	5
Kimi-VL-A3B	✔	63.6	57.8	44.0	45.1	74.9	79.5	82.9	31.6	51.3	84.0	7.4	76.5	83.2	25.4	57.8	6
Qwen3-VL-235B-A22B	✔	66.5	61.4	47.7	46.9	74.8	76.9	94.6	40.4	54.2	92.8	17.2	75.5	94.1	34.5	59.7	11
Kimi-K2.5	✔	69.1	63.6	49.8	51.1	70.8	72.2	96.5	39.4	57.2	95.4	24.1	71.3	96.3	35.5	60.8	12
InternVL-3.5-38B		68.1	63.3	49.1	50.7	75.3	77.1	83.3	26.7	56.6	82.8	6.9	76.0	83.2	21.6	61.5	16
Qwen3-VL-32B	✔	69.2	63.9	50.1	49.4	75.9	76.9	84.3	27.6	56.7	86.0	9.7	76.3	84.7	23.0	61.7	18
GLM-5V	✔	72.7	67.0	53.2	54.3	70.0	71.8	97.4	36.9	60.5	96.9	26.3	70.6	97.3	34.2	63.1	20
GLM-4.6V	✔	67.8	63.2	49.5	49.2	82.6	86.2	93.8	37.7	56.1	91.8	7.9	83.9	93.3	30.1	63.2	22
LLaVA-NeXT-Video-32B		71.2	65.8	51.8	52.4	76.3	78.2	83.7	24.7	59.0	81.1	4.1	77.0	83.0	19.4	63.6	24
Seed-1.5-VL	✔	70.8	65.7	51.2	51.4	79.6	81.6	90.1	31.6	58.4	90.7	10.5	80.3	90.2	26.3	64.0	25
Intern-S1-Pro	✔	73.4	68.6	55.1	55.7	79.6	80.9	93.8	31.8	62.0	94.5	14.4	80.0	94.0	27.4	66.6	26
InternVL-3-78B	✔	81.5	77.3	62.9	62.6	62.7	63.9	87.0	17.2	69.7	69.3	6.2	63.1	82.5	14.4	68.1	27
InternVL-3-8B	✔	74.3	69.9	55.3	56.2	87.1	88.7	88.3	25.6	62.7	90.9	3.2	87.7	88.9	19.9	69.0	28
Qwen2.5-VL-32B		77.8	72.4	58.3	58.0	80.2	81.3	87.9	22.6	65.2	86.9	6.3	80.5	87.6	18.5	69.1	29
MiMo-VL-7B-RL	✔	81.6	76.5	61.9	62.2	77.2	79.4	89.5	20.3	69.2	89.0	11.0	78.0	89.4	18.0	71.4	30
Intern-S1-mini	✔	87.3	82.2	68.2	69.3	77.7	78.3	87.1	11.5	75.5	84.2	6.3	77.9	86.3	10.2	76.1	31
Qwen2.5-VL-7B		86.9	81.2	67.6	67.5	82.0	83.6	87.0	12.5	74.5	89.2	6.7	82.5	87.5	11.0	76.5	32
Proprietary LMMs																	
GPT-5.4	✔	68.2	62.8	49.2	49.5	58.9	59.6	98.7	42.6	56.1	98.5	39.4	59.1	98.6	41.7	56.9	3
Gemini-3.1-Pro	✔	68.4	62.9	48.6	49.1	59.2	63.8	99.3	43.4	55.9	99.1	38.3	60.9	99.2	42.1	57.2	4
o1	✔	69.8	64.2	50.8	49.7	61.5	62.3	90.3	33.1	57.2	91.7	30.0	61.8	90.6	32.3	58.4	7
GPT-5.2	✔	69.5	64.0	49.6	50.6	61.8	63.4	94.6	37.5	57.1	95.2	32.9	62.3	94.7	36.3	58.4	7
Gemini-2.5-Pro	✔	68.5	63.5	49.5	51.2	61.1	65.6	92.2	35.3	57.0	95.2	32.5	62.7	93.0	34.5	58.4	7
Gemini-3-Pro	✔	68.2	63.0	49.9	49.5	65.2	66.5	96.0	39.7	56.3	97.1	31.5	65.7	96.3	37.6	58.7	10
Claude-Opus-4.5	✔	72.0	67.4	53.4	53.4	62.5	63.2	87.6	27.3	60.2	89.0	26.2	62.7	87.9	27.1	60.9	13
Claude-Sonnet-4.6	✔	71.0	66.5	51.9	52.9	65.3	66.9	92.3	33.0	59.3	94.6	28.7	65.9	92.9	31.9	61.0	14
Claude-Opus-4.6	✔	71.2	66.8	52.7	51.8	66.3	68.1	91.9	32.7	59.3	89.3	22.4	66.9	91.3	30.0	61.2	15
Doubao-Seed-1.8	✔	72.8	67.5	53.8	53.3	64.0	65.3	87.0	26.6	60.5	87.2	22.7	64.5	87.1	25.6	61.5	16
GPT-4o	✔	72.6	68.0	53.7	55.4	62.8	63.5	89.4	28.1	61.2	89.1	26.1	63.1	89.3	27.6	61.7	18
o3	✔	73.6	68.2	55.1	54.8	67.3	68.1	89.3	27.7	61.6	89.5	21.9	67.6	89.4	26.3	63.1	20
Grok-3	✔	74.5	70.3	56.3	56.3	64.1	65.5	90.2	27.1	63.1	89.6	25.0	64.6	90.1	26.6	63.5	23
Baseline																	
Human-Best	–	5.6	4.8	2.1	2.0	19.2	24.1	98.6	95.3	3.4	98.8	77.8	21.0	98.7	90.8	7.8	–
Human-Worst	–	8.1	7.4	4.3	4.5	24.1	28.0	92.3	86.4	5.8	92.5	67.0	25.5	92.3	81.5	10.8	–
Vision-Blind†	–	52.1	47.9	34.0	33.7	54.5	56.6	62.1	21.4	40.6	63.4	8.2	55.3	62.4	18.1	44.4	–
Random	–	30.0	28.0	21.8	22.4	23.9	27.1	25.0	0.0	25.0	25.0	0.0	25.0	25.0	0.0	25.0	–

Table 2: Robustness results across 32 LMMs and baselines. The six per-type columns report Robustness Drop (RD) for Information Error (IE: MFE, CFE, VOL, and CMC) and Information Loss (IL: CFL and CIL). Under each dimension, A, R, and D denote reference accuracy ($\mathrm{Acc}_{\mathrm{ref}}$), error-version rejection rate ($\mathrm{Rej}_{\mathrm{err}}$), and robustness drop ($\mathrm{RD} = \mathrm{Acc}_{\mathrm{ref}} - \mathrm{Rej}_{\mathrm{err}}$), respectively. All values are percentages; lower RD indicates better robustness. †Text-only input (no video frames).

Evaluation Protocol. Each model runs under its official video configuration (full frame budgets in Appendix) with temperature 0, except o1 and o3, which use their minimum supported temperature of 1. The prompt presents only the question stem and four options (A–D) with an instruction that models may output `unanswerable` when no option is correct. This probes whether models can appropriately exercise conditional rejection when degradation renders all options invalid as examined in the prompt-strategy ablation.

Rejection Determination. A response is a rejection if it explicitly declines all four options. We use a conservative two stage procedure. Stage 1 applies keyword matching (see Appendix) flagging outputs with uncertainty phrases and no option letter present. Stage 2 submits ambiguous cases to GPT-4o for binary classification. GPT-4o responses are judged only by Stage 1 to avoid self evaluation bias. Empty or malformed outputs are treated as incorrect answers.

Baselines. **Random** yields 25% reference accuracy and near zero rejection ($\mathrm{RD} \approx 25\%$). **Vision Blind** runs Gemini-2.5-Pro on text-only input comprising the question stem, four answer options, and the ASR transcript, without video frames. **Human** uses three domain experts on a stratified

subset of 300 questions, with each expert independently providing both answer selection and answerability judgment. Human-Best and Human-Worst denote the individual experts with the highest and lowest aggregate performance, respectively, rather than per-question oracle bounds.

Metrics. The primary metric is **Robustness Drop** ($\mathrm{RD} = \mathrm{Acc}_{\mathrm{ref}} - \mathrm{Rej}_{\mathrm{err}}$). RD is reported only when $\mathrm{Acc}_{\mathrm{ref}}$ exceeds the random baseline by a reliable margin. Supporting metrics include $\mathrm{Acc}_{\mathrm{ref}}$ (clean accuracy) and $\mathrm{Rej}_{\mathrm{err}}$ (error-version rejection). We estimate 95% confidence intervals by paired question-level bootstrap resampling and use paired bootstrap tests for pre-specified comparisons; the complete interval and test results are reported in the technical supplement.

4.2 Main Results and Analysis

Overall Robustness Performance. Every model yields $\mathrm{RD} > 0$ (Table 2), confirming a universal robustness gap. Proprietary models average higher $\mathrm{Acc}_{\mathrm{ref}}$ and $\mathrm{Rej}_{\mathrm{err}}$, but even the best LMM rejection rate (GPT-5.4, 41.7%) remains far below Human-Worst. Per-model rejection on answerable reference versions is reported in the technical supplement to distinguish degradation awareness from indiscriminate abstention.

Degradation Dimension Analysis: IE versus IL. Across all models, IL types elicit lower $\mathrm{Rej}_{\mathrm{err}}$ than IE types: 18.4% on IL versus 30.5% on IE. Consequently, IL yields substantially higher RD than IE. IE types inject erroneous content that models can occasionally detect through domain priors or cross-modal comparison (e.g., physics violations in VOL). IL removes information from the question stem (CFL) or video (CIL), leaving all four options plausible and providing no detectable anomaly. Even on IE, $\mathrm{Rej}_{\mathrm{err}}$ falls far below human performance, indicating that cross-modal verification remains a consistent weakness.

Per-Type Robustness Breakdown. Table 2 shows the per-type RD averaged across all 32 LMMs, in descending order: CIL$>$ MFE $>$ CFL$>$ CFE $>$ CMC $>$ VOL. IL types occupy two of the top three positions, consistent with the higher overall IL RD. CIL ranks highest: blank-frame substitution is visually inconspicuous, providing no signal that content is missing, and contains the smallest subset ($N = 273$). MFE ranks second: numerical deviations in experimental parameters (e.g., reagent concentrations, temperature values) are the hardest IE type to detect, as they require precise quantitative verification against the video content without any visual anomaly to cue inspection. CFL follows: the absent stem field leaves all options plausible, and models lack a mechanism to assess answerability. CFE provides stronger detection cues through entity substitutions, yielding moderately lower RD. CMC and VOL show the lowest RD among all types: VOL are partially flagged by domain priors, while cross-modal contradictions in CMC occasionally trigger verification, though detection remains weak in absolute terms.

Model Organization and Reasoning Strategy. Proprietary models achieve a lower mean RD than open-source models, driven primarily by the IL dimension where the gap is 12.6 points compared to only 1.4 points on IE. This asymmetry is notable: IL degradation removes information, providing no factual signal for domain knowledge to exploit; the proprietary advantage on IL may therefore reflect differences in calibration or a stronger prior that SOMETHING IS MISSING rather than superior content analysis. Models with explicit chain-of-thought reasoning (Think, 25 of 32 LMMs) exhibit a negligible RD advantage over non-thinking models. The best-performing model overall, Qwen2.5-VL-72B, lacks explicit thinking, while several thinking-capable models rank near the bottom, indicating that post-training for reasoning does not automatically confer robustness to visual degradation. A possible explanation is that explicit chain-of-thought reasoning elaborates step-by-step arguments for a particular answer choice, which may increase commitment to that choice and reduce the likelihood of stepping back to question whether the input itself is reliable.

Question Type Analysis. The SD–EC comparison reveals a model-dependent rather than universal question-type effect (Table 3). For GPT-5.2, EC questions yield lower $\mathrm{Acc}_{\mathrm{ref}}$, lower $\mathrm{Rej}_{\mathrm{err}}$, and a slightly higher RD. Qwen2.5-VL-72B shows a similarly small increase in EC RD, whereas GPT-4o, GLM-4.6V, and InternVL-3.5-38B show lower EC RD. Thus, robustness gap persists across both question types, but its magnitude is not solely determined by SD–EC distinction.

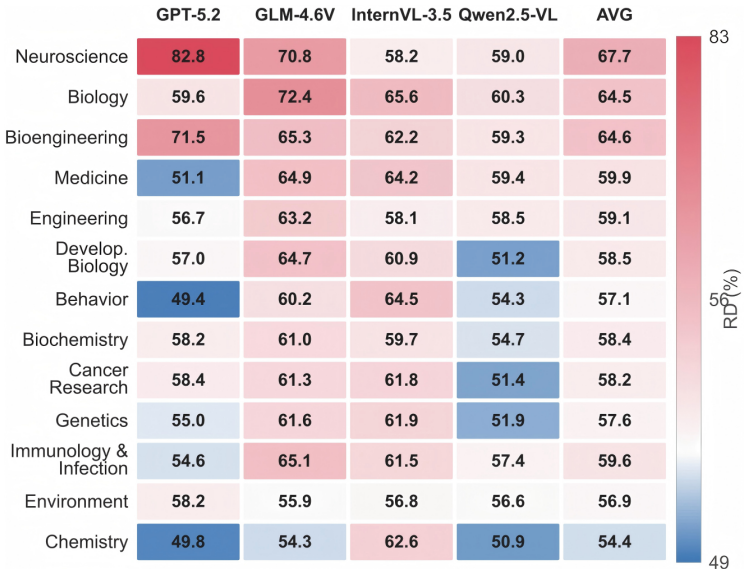


Figure 3: Per-discipline Robustness Drop for four representative models spanning open-source and proprietary families. Black line: cross-model average.

Cross-Discipline Robustness. RD is high across all 13 disciplines, showing that the robustness deficit spans all evaluated domains (Figure 3). Specialized disciplines show relatively higher IE RD than foundational disciplines. Although SD/EC ratios and degradation-type distributions are balanced within $\pm 5\%$, video duration, question difficulty, and discipline-specific sample sizes may also contribute to the remaining variation.

Model Scale and Architecture. Among open-source models, $\mathrm{Acc}_{\mathrm{ref}}$ shows no statistically significant correlation with scale ($r = +0.49$, $p = 0.13$). Within-family trends are inconsistent: Qwen2.5-VL 72B achieves lower $\mathrm{Acc}_{\mathrm{ref}}$ than its 7B and 32B counterparts, and InternVL-3 drops from 88.9% (8B) to 82.5% (78B). $\mathrm{Rej}_{\mathrm{err}}$ shows a moderate positive correlation with scale ($r = +0.67$, $p = 0.03$), but the direction is not uniform across families: Qwen2.5-VL rejection rises from 11.0% (7B) to 28.9% (72B), while InternVL-3 rejection declines from 19.9% (8B) to 14.4% (78B). The mixed family-level patterns suggest that parameter count alone is an insufficient predictor of degradation awareness. Factors such as training data composition, post-training procedures, or architectural design choices may exert greater influence. The opposing within-family trends show that the aggregate association does not constitute a consistent scaling law. Model size therefore cannot serve as a reliable proxy for degradation robustness across model families.

4.3 Ablation Studies

Vision-Blind Validation. We evaluate Gemini-2.5-Pro on text-only input comprising the question stem, four answer options, and the ASR transcript, without video frames. Under this vision-blind condition, $\mathrm{Acc}_{\mathrm{ref}}$ drops to 62.41% from 92.99%. Under the vision-blind condition, the model continues to commit to an option on most degraded inputs rather than identifying the missing visual evidence. This result shows that the textual stream alone does not provide sufficient evidence for reliable degradation detection and motivates comparison with the full video condition.

Paired Behavioral Analysis. The paired design of RoSci-Video enables per-question behavioral analysis. R^- dominates the paired outcomes (63.9% on average), whereas W^+ is only 0.5% (Table 5); thus, RD is driven primarily by models that answer the reference correctly but commit under degradation. R^+ accounts for only 24.5%, showing that explicit rejection instructions rarely trigger appropriate abstention. These results of the paired subset in Technical Supplement.

Model	SD			EC			ΔRD
	$\mathrm{Acc}_{\mathrm{ref}}$	$\mathrm{Rej}_{\mathrm{err}}$	RD	$\mathrm{Acc}_{\mathrm{ref}}$	$\mathrm{Rej}_{\mathrm{err}}$	RD	
GPT-5.2	94.8	54.2	40.6	88.7	46.5	42.2	+1.6
GPT-4o	90.9	32.4	58.5	82.2	25.5	56.8	−1.8
GLM-4.6V	92.2	24.5	67.7	89.4	28.9	60.6	−7.1
InternVL-3.5-38B	87.3	0.3	87.1	79.9	0.3	79.6	−7.4
Qwen2.5-VL-72B	92.0	22.1	69.8	87.6	16.3	71.3	+1.5

Table 3: Robustness metrics (%) by question type on the paired evaluation subset. $\Delta\mathrm{RD} = \mathrm{RD}_{\mathrm{EC}} - \mathrm{RD}_{\mathrm{SD}}$; positive values indicate a larger robustness drop on EC questions.

Model	Overall	IE	IL	$\mathbf{R}^-$
GPT-5.2	46.2	45.9	47.1	40.1
GPT-4o	39.7	43.3	29.4	3.7

Table 4: Answer Consistency Rate (ACR, %). Overall ACR is the fraction of committed error-version responses that match the reference-version answer. R^- restricts ACR to reference-correct, error-committed cases.

Model	Paired decomposition					Reject-instruction		
	R^+	R^-	W^+	W^-	**RD**	Rej_A	Rej_B	**Δ**
GPT-5.2	49.6	42.1	0.6	7.7	41.4	36.3	17.6	18.7
GPT-4o	28.2	58.3	0.7	12.8	57.6	27.6	8.3	19.3
GLM-4.6V	26.0	64.8	0.7	8.5	64.2	30.1	15.3	14.8
InternVL-3.5-38B	0.3	83.2	0.0	16.5	83.2	21.6	7.4	14.2
Qwen2.5-VL-72B	18.4	71.3	0.7	9.6	70.6	28.9	11.5	17.4
AVG	24.5	63.9	0.5	11.0	63.4	28.9	12.0	16.9

Table 5: Paired outcome decomposition (left) and rejection-instruction ablation (right) . (left) The outcome decomposition uses the paired evaluation subset, whereas the prompt ablation uses the full ablation evaluation. R^+: reference correct and error rejected. R^-: reference correct and error committed. W^+: reference incorrect and error rejected. W^-: reference incorrect and error committed. (right) Method A uses an explicit `unanswerable` instruction and Method B uses no rejection instruction, where $\Delta = \mathrm{Rej}_A - \mathrm{Rej}_B$.

Committed-Answer Consistency. ACR distinguishes failures that preserve the clean-input answer from those that change the committed answer under degradation. Table 4 reports the two models selected for the main-text comparison; expanded ACR results for all models with option-level response logs are provided in the technical supplement. GPT-5.2 and GPT-4o achieve overall ACR values of 46.2% and 39.7%, respectively, both above the 25% random baseline (Table 4). However, their capability-present failures differ sharply. GPT-5.2 retains an R^- ACR of 40.1%, indicating that many failures reproduce the clean-input answer despite degradation, whereas GPT-4o reaches only 3.7%, indicating that degradation more often shifts its committed answer. For these two models, positive RD can therefore arise from either preserving a clean-input answer despite degraded evidence or changing the committed answer under degradation.

Prompt Strategy Comparison. Method A explicitly permits `unanswerable`, whereas Method B omits the rejection instruction. Removing this instruction reduces $\mathrm{Rej}_{\mathrm{err}}$ from 28.9% to 12.0% (Table 5), with substantial variation across models. Method A has only a small effect on $\mathrm{Acc}_{\mathrm{ref}}$ in the corresponding clean-input control (full per-model results are reported in the technical supplement), indicating that its rejection gain is concentrated on degraded inputs rather than indiscriminate refusal. The remaining gap under Method A therefore cannot be explained by the absence of an explicit rejection option alone. Together, the paired decomposition, ACR analysis, and prompt ablation separate the availability of a rejection response from a model's ability to recognize when that response is warranted. Prompting increases rejection, but the remaining R^- cases show that degradation detection is the central bottleneck.

5 Conclusion

RoSci-Video constitutes a systematic benchmark that exposes a previously overlooked failure mode of LMMs in scientific understanding and reasoning: the inability to detect information degradation and appropriately refuse to answer. Our evaluation shows that even models with strong clean video accuracy consistently fail to recognize degraded inputs. Within-family robustness trends are inconsistent, indicating that scale alone does not reliably predict degradation robustness. This dissociation identifies degradation awareness as a capability that requires explicit evaluation and design. RoSci-Video provides the first structured testbed for measuring this capability and establishes Robustness Drop as a principled metric for calibration under semantic corruption. We hope this benchmark spurs research on epistemic uncertainty estimation and refusal mechanisms, and encourages LMMs to handle scientific discovery tasks more reliably and rationally.

AI USE STATEMENT

We used generative AI tools solely to assist with LaTeX formatting and figure and table layout. The authors reviewed the resulting changes and take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human-subject research. The perturbed inputs in RoSci-Video are intended solely for robustness evaluation and should not be treated as reliable scientific information. Any reuse or redistribution of source videos should respect applicable licenses and the rights of content owners.

REPRODUCIBILITY STATEMENT

The main text and appendix describe the benchmark construction, degradation types, evaluation protocols, and metric definitions to support reproducibility.

REFERENCES

Anthropic. Claude Opus 4.5 system card. https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf, 2025. Accessed: 2026-07-28.

Anthropic. Claude Opus 4.6 system card. https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf, 2026a. Accessed: 2026-07-28.

Anthropic. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, 2026b. Accessed: 2026-07-28.

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, et al. Intern-S1: A scientific multimodal foundation model, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025b. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report, 2025c. URL https://arxiv.org/abs/2502.13923.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. MicroVQA: A multimodal reasoning benchmark for Microscopy-Based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19552–19564, June 2025.

Bytedance Seed. Seed1.8 Model Card: Towards generalized real-world agency, 2026. URL https://arxiv.org/abs/2603.20633.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Julian Eisenschlos, Hernán Maina, Guido Ivetta, and Luciana Benotti. Selectively answering visual questions. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4219–4229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.250. URL https://aclanthology.org/2024.findings-acl.250/.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 24108–24118, June 2025.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026a.

GLM-V Team. GLM-5V-Turbo: Toward a native foundation model for multimodal agents, 2026b.

Google DeepMind. Gemini 3 Pro model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf`, 2025. Accessed: 2026-07-28.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-VL technical report, 2025.

Nidhal Jegham, Marwan F. Abdelatti, and Abdeltawab M. Hendawi. Visual reasoning consistency and robustness analysis of multimodal LLMs. *Pattern Recognition*, 172:112765, 2026. doi: 10.1016/j.patcog.2025.112765.

Kimi Team. Kimi-VL technical report, 2025.

Kimi Team. Kimi K2.5: Visual agentic intelligence, 2026.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177.

Arka Mukherjee and Shreya Ghosh. mmJEE-eval: A bilingual multimodal benchmark for evaluating scientific reasoning in vision-language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2268–2290, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-ijcnlp.140.

OpenAI. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

OpenAI. Update to GPT-5 system card: GPT-5.2. `https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf`, 2025a. Accessed: 2026-07-28.

OpenAI. Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025b. Accessed: 2026-07-28.

OpenAI. Openai o1 system card, 2026a. URL https://arxiv.org/abs/2412.16720.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026b. Accessed: 2026-07-28.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9014–9023, October 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. Accessed: 2026-07-28.

The Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026. Accessed: 2026-07-28.

Junying Wang, Zicheng Zhang, Yijin Guo, Farong Wen, Ye Shen, Yingji Liang, Yalun Wu, Wenzhe Li, Chunyi Li, Zijian Chen, Qi Jia, and Guangtao Zhai. The ever-evolving science exam, 2025a. URL https://arxiv.org/abs/2507.16514.

Junying Wang, Zicheng Zhang, Ye Shen, Yalun Wu, Yingji Liang, Yijin Guo, Farong Wen, Wenzhe Li, Xuezhi Zhao, Qi Jia, and Guangtao Zhai. Q-mirror: Unlocking the multi-modal potential of scientific text-only qa pairs, 2025b. URL https://arxiv.org/abs/2509.24297.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025c.

Jingxuan Wei, Nan Xu, Guiyong Chang, Yin Luo, Bihui Yu, and Ruifeng Guo. mChartQA and mChartQABench: A multimodal-only solution for complex chart question-answering. *Pattern Recognition*, 172:112348, 2026. doi: 10.1016/j.patcog.2025.112348.

xAI. Grok 3 Beta — the age of reasoning agents. https://x.ai/news/grok-3, 2025. Accessed: 2026-07-28.

Xiaomi LLM-Core Team. MiMo-VL technical report, 2025.

Tianchi Xie, Minzhi Lin, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. InfoChartQA: A benchmark for multimodal question answering on infographic charts, 2025. URL https://arxiv.org/abs/2505.19028.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, Manabu Okumura, and Yi Wang. ExpVid: A benchmark for experiment video understanding & reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bj0ahFM4P0.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (MMIR): A new benchmark for multimodal reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18829–18845, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.964.

Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. MM-InstructEval: Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. *Information Fusion*, 122:103204, 2025. doi: 10.1016/j.inffus.2025.103204.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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, June 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Z.AI. GLM-4.6V: Open source multimodal models with native tool use. https://z.ai/blog/glm-4.6v, 2025. Accessed: 2026-07-28.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024. Accessed: 2026-07-28.

Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, Bowen Yan, Ziheng Jia, Jiahao Xiao, Yuan Tian, Xiangyang Zhu, Kaiwei Zhang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Qi Jia, and Guangtao Zhai. Aibench: Towards trustworthy evaluation under the 45° law. *Displays*, pp. 103255, 2025a. ISSN 0141-9382. doi: 10.1016/j.displa.2025.103255.

Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, Shengyuan Ding, Ziheng Jia, Jiahao Xiao, Ye Shen, Yushuo Zheng, Xiaorong Zhu, Yalun Wu, Ziheng Jiao, Wei Sun, Zijian Chen, Kaiwei Zhang, Kang Fu, Yuqin Cao, Ming Hu, Yue Zhou, Xuemei Zhou, Juntai Cao, Wei Zhou, Jinyu Cao, Ronghui Li, Donghao Zhou, Yuan Tian, Xiangyang Zhu, Chunyi Li, Haoning Wu, Xiaohong Liu, Junjun He, Yu Zhou, Hui Liu, Lin Zhang, Zesheng Wang, Huiyu Duan, Yingjie Zhou, Xiongkuo Min, Qi Jia, Dongzhan Zhou, Wenlong Zhang, Jiezhang Cao, Xue Yang, Junzhi Yu, Songyang Zhang, Haodong Duan, and Guangtao Zhai. Large multimodal models evaluation: A survey. *SCIENCE CHINA Information Sciences*, 68(12):221301, 2025b. doi: https://doi.org/10.1007/s11432-025-4676-4.

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

Yicheng Zou, Dongsheng Zhu, Lin Zhu, Tong Zhu, Yunhua Zhou, Peiheng Zhou, Xinyu Zhou, Dongzhan Zhou, Zhiwang Zhou, Yuhao Zhou, et al. Intern-S1-Pro: Scientific multimodal foundation model at trillion scale, 2026.

A TECHNICAL APPENDICES

A.1 MODEL EVALUATION CONFIGURATION

Table 6 summarizes the evaluation configuration for all 32 LMMs. All models process uniformly sampled video frames resized to 512×512. Frames are extracted at 1 fps; when the video frame count exceeds the budget, a uniform subset is taken. All models run with temperature 0, except o1 and o3 which use the minimum supported temperature of 1. Open-source models use official checkpoints with greedy decoding (non-thinking) or default reasoning budget (thinking models). Gemini-3.1-Pro and Gemini-3-Pro process frames at 1 fps through the native video API, effectively seeing all available frames (average video length: $\sim$503 s, yielding $\sim$500 frames per sample). Grok-3's native video API does not expose a configurable frame budget; its internal frame handling is provider-defined.

Rejection Keywords. We apply a two-stage rejection detection pipeline. Stage 1 flags a response as a rejection if it contains any of the following keywords or phrases: 'unanswerable', 'cannot be answered', 'cannot determine', 'not answerable', 'insufficient information', 'none of the above', 'no correct option', 'unable to', 'impossible to determine', 'cannot be concluded', or semantically equivalent expressions in the target language, *and* no option letter (A, B, C, D) appears as a standalone selection. Stage 2 submits residual ambiguous cases (especially responses that contain both rejection language and option references) to GPT-4o for binary classification. When GPT-4o is itself an evaluated model, its responses are judged only by Stage 1 to avoid self-evaluation bias. Empty or malformed responses, as well as responses that consist exclusively of punctuation or whitespace, are treated as incorrect answers.

A.2 STATISTICAL INFERENCE

All uncertainty estimates use paired question-level bootstrap resampling, with the reference and degraded response for each question kept in the same resampled unit. Unless otherwise stated, confidence intervals are two-sided 95% percentile intervals and paired bootstrap tests use the same resampled question indices. At the point-estimate level, all 32 LMMs have positive RD, ranging from 55.7% to 76.5%. All 32 lower confidence bounds are also above zero; their lower endpoints range from 53.6% to 74.9%. The observed model-level tests are all below 0.001. Thus, the full-benchmark inference supports a systematic robustness gap rather than a gap confined to a small subset of models.

Summary of pre-specified comparisons. Information loss produces lower rejection than information error (18.4% versus 30.5%, a difference of -12.1 percentage points). Removing the rejection instruction decreases degraded-input rejection from 28.9% to 12.0%, a difference of 16.9 percentage points. These results show that information loss suppresses rejection, whereas explicitly allowing rejection substantially increases rejection on degraded inputs.

Reference-Version Rejection. The final column of Table 8 reports rejection on clean reference versions, which are answerable by construction. This quantity measures indiscriminate abstention rather than degradation awareness. The rate is $100 \times \text{Rejected}/3000$. Reference-version rejection ranges from 0.1% to 1.9% across models, with Qwen2.5-VL-7B showing the highest rate and GPT-5.4 and Gemini-3.1-Pro tied for the lowest. These values quantify the extent to which a model abstains even when the clean evidence supports one of the four options; clean-input rejection remains far below degraded-input rejection.

A.3 PAIRWISE ANALYSIS OF REPRESENTATIVE MODELS

We conduct detailed per-question paired analysis on four focal LMMs (GPT-5.2, GLM-4.6V, InternVL-3.5-38B, and Qwen2.5-VL-72B) that span a range of robustness characteristics; GPT-4o is additionally included as a proprietary API baseline where comparable response logs are available. The paired subset was formed by uniformly sampling 60% of the full benchmark, yielding 1,800 of the 3,000 paired questions. Sampling was performed before inspecting model outcomes and used the same question identifiers for every included model. Consequently, differences across models are computed on an identical set of reference–error pairs. Because the subset was selected by uniform random sampling rather than outcome-dependent filtering, it is not constructed to favor any model

Model	Org	Think	Frames (max)	Temp.
Open-source LMMs (19 models)				
Qwen2.5-VL-7B	Alibaba		128	0
Qwen2.5-VL-32B	Alibaba		768	0
Qwen2.5-VL-72B	Alibaba		128	0
Qwen3-VL-32B	Alibaba/Qwen	✔	128	0
Qwen3.5-35B-A3B	Alibaba/Qwen		128	0
Qwen3-VL-235B-A22B	Alibaba/Qwen	✔	128	0
InternVL-3-8B	AILab	✔	256	0
InternVL-3-78B	AILab	✔	256	0
InternVL-3.5-38B	AILab		256	0
Intern-S1-mini	AILab	✔	128	0
Intern-S1-Pro	AILab	✔	128	0
LLaVA-NeXT-Video-32B	ByteDance		32	0
GLM-4.5V	ZhipuAI	✔	256	0
GLM-4.6V	ZhipuAI	✔	256	0
GLM-5V	ZhipuAI	✔	256	0
Kimi-VL-A3B	Moonshot AI	✔	256	0
Kimi-K2.5	Moonshot AI	✔	256	0
MiMo-VL-7B-RL	Xiaomi	✔	512	0
Seed-1.5-VL	ByteDance	✔	256	0
Proprietary LMMs (13 models)				
GPT-5.4	OpenAI	✔	128	0
Gemini-3.1-Pro	Google	✔	1 fps	0
GPT-5.2	OpenAI	✔	128	0
GPT-4o	OpenAI	✔	256	0
o1	OpenAI	✔	128	1†
o3	OpenAI	✔	128	1†
Gemini-2.5-Pro	Google	✔	128	0
Gemini-3-Pro	Google	✔	1 fps	0
Claude-Sonnet-4.6	Anthropic	✔	96	0
Claude-Opus-4.5	Anthropic	✔	96	0
Claude-Opus-4.6	Anthropic	✔	96	0
Grok-3	xAI	✔	–‡	0
Doubao-Seed-1.8	ByteDance	✔	256	0

Table 6: Evaluation configuration for all 32 LMMs. 'Frames': maximum video frames per sample; '1 fps' denotes native video input at 1 fps. Temperature: 0 for all models except o1 and o3 ($T = 1$, minimum for OpenAI reasoning models). † denotes the minimum supported temperature for o1 and o3; ‡ denotes provider-defined frame handling for Grok-3.

or degradation type. For each included model, both reference-version and error-version responses are available, enabling direct computation of $R^+/R^-/W^+/W^-$, RD, and ACR at the question level. Because this subset differs from the full benchmark, its absolute RD values should not be compared directly with the full-dataset ranking in the main results table.

A.3.1 Overall Robustness Comparison

Table 7 summarizes performance on the fixed paired evaluation subset.

Key observations. (i) Every model exhibits a positive RD on this paired subset, confirming a robustness gap. (ii) InternVL-3.5-38B achieves near-zero rejection (0.3%) and the highest RD on this subset, demonstrating that high reference accuracy without rejection ability yields the weakest

Model	Overall			IE			IL		
	$\mathrm{Acc}_{\mathrm{ref}}$	$\mathrm{Rej}_{\mathrm{err}}$	RD	$\mathrm{Acc}_{\mathrm{ref}}$	$\mathrm{Rej}_{\mathrm{err}}$	RD	$\mathrm{Acc}_{\mathrm{ref}}$	$\mathrm{Rej}_{\mathrm{err}}$	RD
GPT-5.2	91.7%	50.2%	41.4	92.4%	52.0%	40.4	89.5%	45.1%	44.3
GPT-4o	86.5%	28.9%	57.6	87.8%	30.6%	57.2	82.7%	23.9%	58.8
GLM-4.6V	90.8%	26.7%	64.2	91.3%	26.4%	64.9	89.3%	27.4%	62.0
InternVL-3.5-38B	83.5%	0.3%	83.2	84.6%	0.2%	84.4	80.3%	0.4%	79.9
Qwen2.5-VL-72B	89.7%	19.1%	70.6	90.8%	20.1%	70.7	86.4%	16.1%	70.3
AVG	88.4%	25.0%	63.4	89.4%	25.9%	63.5	85.6%	22.6%	63.1

Table 7: Overall robustness metrics for four focal models and the additional GPT-4o baseline on the paired evaluation subset. $\mathrm{RD} = \mathrm{Acc}_{\mathrm{ref}} - \mathrm{Rej}_{\mathrm{err}}$. Every model exhibits positive RD; InternVL-3.5-38B shows near-zero rejection (0.3%) and the weakest robustness on this subset ($\mathrm{RD} = 83.2$), demonstrating that high accuracy without degradation awareness yields the largest robustness gap. Values may not subtract exactly from the displayed rounded percentages.

Model	RD (%)	95% CI	p vs. 0	Reference rejection (%)
Qwen2.5-VL-72B	55.7	53.6–57.8	0.0003	0.9
GLM-4.5V	56.8	54.9–58.9	0.0007	0.8
Qwen3.5-35B-A3B	57.4	55.3–59.3	0.0002	0.8
Kimi-VL-A3B	57.8	55.7–59.9	0.0009	1.2
Qwen3-VL-235B-A22B	59.7	57.7–61.5	0.0004	0.5
Kimi-K2.5	60.8	59.0–62.6	0.0006	0.4
InternVL-3.5-38B	61.5	59.6–63.6	0.0001	1.2
Qwen3-VL-32B	61.7	59.7–63.7	0.0008	1.0
GLM-5V	63.1	61.3–64.9	0.0005	0.3
GLM-4.6V	63.2	61.3–65.1	0.0002	0.5
LLaVA-NeXT-Video-32B	63.6	61.6–65.6	0.0007	1.5
Seed-1.5-VL	64.0	62.0–65.8	0.0003	0.8
Intern-S1-Pro	66.6	64.8–68.4	0.0009	0.4
InternVL-3-78B	68.1	66.2–70.0	0.0005	0.8
InternVL-3-8B	69.0	67.2–70.8	0.0001	1.5
Qwen2.5-VL-32B	69.1	67.3–70.9	0.0006	1.3
MiMo-VL-7B-RL	71.4	69.6–73.2	0.0004	0.6
Intern-S1-mini	76.1	74.5–77.7	0.0008	1.0
Qwen2.5-VL-7B	76.5	74.9–78.1	0.0002	1.9
GPT-5.4	56.9	55.1–58.7	0.0007	0.1
Gemini-3.1-Pro	57.2	55.3–58.9	0.0003	0.1
o1	58.4	56.3–60.3	0.0009	0.3
GPT-5.2	58.4	56.5–60.3	0.0005	0.3
Gemini-2.5-Pro	58.4	56.6–60.4	0.0001	0.3
Gemini-3-Pro	58.7	56.8–60.6	0.0006	0.2
Claude-Opus-4.5	60.9	58.8–62.8	0.0004	0.6
Claude-Sonnet-4.6	61.0	59.1–62.9	0.0002	0.4
Claude-Opus-4.6	61.2	59.4–63.2	0.0008	0.4
Doubao-Seed-1.8	61.5	59.5–63.5	0.0003	0.7
GPT-4o	61.7	59.8–63.6	0.0007	0.7
o3	63.1	61.2–65.0	0.0005	0.2
Grok-3	63.5	61.6–65.4	0.0009	0.7

Table 8: Full-benchmark robustness-drop estimates, 95% percentile bootstrap confidence intervals, model-level tests against $\mathrm{RD} = 0$, and rejection rates on answerable reference versions. All values are percentages except the reported p values. The model-level p values are reported without multiplicity correction as descriptive tests for each model.

robustness. (iii) GPT-5.2 achieves the lowest RD on this subset (41.4) through balanced accuracy and rejection. (iv) The IL dimension produces higher RD than IE for GPT-5.2 (44.3 vs. 40.4), consistent with the finding that missing information is harder to detect than erroneous content.

Model	Rej_A (%)	Rej_B (%)	Δ (%)
GPT-5.2	36.3	17.6	18.7
GPT-4o	27.6	8.3	19.3
GLM-4.6V	30.1	15.3	14.8
InternVL-3.5-38B	21.6	7.4	14.2
Qwen2.5-VL-72B	28.9	11.5	17.4
AVG	28.9	12.0	16.9

Table 9: Rejection-instruction ablation on the full ablation evaluation. Method A explicitly permits `unanswerable`; Method B omits the rejection instruction. All values are percentages, and Δ is the absolute increase in rejection under Method A.

Model	$\mathrm{Acc}_{\mathrm{ref},A}$ (%)	$\mathrm{Acc}_{\mathrm{ref},B}$ (%)	Δ (pp)	N
GPT-5.2	94.7	94.4	+0.3	3000
GPT-4o	89.4	89.2	+0.2	3000
GLM-4.6V	93.2	92.8	+0.4	3000
InternVL-3.5-38B	83.1	82.9	+0.2	3000
Qwen2.5-VL-72B	84.8	84.4	+0.4	3000
AVG	89.0	88.8	+0.3	3000

Table 10: Observed clean-input control for the rejection-instruction ablation. Method A includes the rejection instruction and Method B omits it. All five models are evaluated on the fixed set of 3,000 matched reference questions. Differences and averages are computed from the unrounded underlying accuracy values, so they may not equal arithmetic on the displayed one-decimal percentages.

A.3.2 Prompt Strategy Comparison

The rejection-instruction ablation uses the full ablation evaluation rather than the paired subset above. Method A presents the instruction that the model may output `unanswerable` when none of the four options is correct; Method B removes this rejection instruction and retains the ordinary four-option answer format. Table 9 reports the complete per-model rejection results used for the main-text comparison. Here Rej_A and Rej_B are rejection rates on the degraded inputs, and $\Delta = \mathrm{Rej}_A - \mathrm{Rej}_B$.

For the clean-input control, the corresponding accuracy comparison is reported separately because reference versions are answerable and should not be rejected. Method A changes clean-input accuracy only slightly: the mean accuracy is 89.0% versus 88.8% for Method B, an average increase of 0.3 percentage points. Across the five listed models, the observed changes range from 0.1 to 0.5 percentage points.

A.3.3 Per-Model Detailed Breakdown

GPT-5.2. GPT-5.2 achieves the best robustness on this paired subset (RD 41.4) through balanced $\mathrm{Acc}_{\mathrm{ref}}$ (91.7%) and $\mathrm{Rej}_{\mathrm{err}}$ (50.2%). By type, VOL shows the lowest RD (35.9) and CFL the highest (44.3). By dimension, IL RD (44.3) exceeds IE RD (40.4). R^- ACR ranges from 38.5% (CFL) to 41.4% (VOL).

GPT-4o. GPT-4o shows moderate robustness (RD 57.6) with $\mathrm{Acc}_{\mathrm{ref}}$ 86.5% and $\mathrm{Rej}_{\mathrm{err}}$ 28.9%. Per-type RD is relatively uniform (53.5–59.8), and IL RD (58.8) slightly exceeds IE RD (57.2). GPT-4o's low R^- ACR (3.7%) indicates that when it commits on error versions, it rarely reproduces its reference-version answer, suggesting degradation shifts its reasoning.

GLM-4.6V. GLM-4.6V shows RD 64.2 with $\mathrm{Acc}_{\mathrm{ref}}$ 90.8% and $\mathrm{Rej}_{\mathrm{err}}$ 26.7%. Unusually, IL RD (62.0) is lower than IE RD (64.9), unlike the other models.

InternVL-3.5-38B. InternVL-3.5-38B shows the weakest robustness on this paired subset (RD 83.2) due to near-zero rejection (0.3%) despite moderate accuracy (83.5%). This represents the

Model	Overall (%)	IE (%)	IL (%)	R^- (%)
Qwen2.5-VL-72B	42.1	44.7	39.3	35.6
GLM-4.5V	40.5	42.9	36.2	32.4
Qwen3.5-35B-A3B	39.0	41.3	35.7	30.1
Kimi-VL-A3B	38.6	40.8	34.5	29.3
Qwen3-VL-235B-A22B	45.2	47.8	42.6	39.4
Kimi-K2.5	46.5	48.9	43.7	40.2
InternVL-3.5-38B	35.3	37.1	32.8	22.5
Qwen3-VL-32B	38.1	40.6	35.9	30.7
GLM-5V	47.4	49.6	44.3	42.1
GLM-4.6V	44.0	46.7	41.5	38.6
LLaVA-NeXT-Video-32B	36.8	38.5	33.2	26.4
Seed-1.5-VL	43.3	45.9	40.1	35.2
Intern-S1-Pro	46.9	48.4	43.8	41.2
InternVL-3-78B	33.6	35.7	30.3	20.8
InternVL-3-8B	34.2	36.9	31.4	22.1
Qwen2.5-VL-32B	35.5	37.8	32.0	26.0
MiMo-VL-7B-RL	32.4	34.6	29.7	21.3
Intern-S1-mini	30.0	32.2	27.6	18.4
Qwen2.5-VL-7B	29.1	31.5	26.8	15.2
GPT-5.4	47.1	49.3	45.6	42.0
Gemini-3.1-Pro	48.2	50.7	46.1	43.5
o1	44.6	46.3	41.8	37.5
GPT-5.2	46.2	45.9	47.1	40.1
Gemini-2.5-Pro	45.8	47.0	42.4	39.1
Gemini-3-Pro	47.6	49.9	45.0	41.7
Claude-Opus-4.5	40.3	42.8	37.6	33.9
Claude-Sonnet-4.6	44.4	46.6	41.1	37.4
Claude-Opus-4.6	43.1	45.4	40.5	36.3
Doubao-Seed-1.8	39.6	41.9	36.6	31.8
GPT-4o	39.7	43.3	29.4	3.7
o3	42.2	44.9	39.8	35.1
Grok-3	41.0	43.2	38.3	34.0

Table 11: Full-model Answer Consistency Rate (ACR, %). All 32 rows are observed option-level point estimates reported to one decimal place. Overall ACR is computed among committed error-version responses; R^- ACR is restricted to reference-correct, error-committed cases.

extreme case where a model possesses the knowledge to answer correctly but completely lacks degradation awareness. The paired outcome category R^- accounts for 83.2% of all questions.

Qwen2.5-VL-72B. Qwen2.5-VL-72B shows RD 70.6 with $\mathrm{Acc}_{\mathrm{ref}}$ 89.7% and $\mathrm{Rej}_{\mathrm{err}}$ 19.1%. IL RD (70.3) is slightly lower than IE RD (70.7), indicating similar difficulty detecting both degradation categories.