VIDEO-INDEX: A CURATED META-BENCHMARK FOR VIDEO UNDERSTANDING EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A video benchmark should reward the capability it claims to measure, yet models can exploit answer options, question text, or partial visual evidence. We introduce the attack pyramid, five levels of shortcut attacks with increasing access to each item, and audit 115 video benchmarks with it. On 35 benchmarks, attackers that never see a frame approach full-video accuracy. On 51 benchmarks with temporal probes, shuffled frames keep a median 96% of full-video accuracy. Near-duplicate questions make up at least half the items in 63 benchmarks. We screen 505,518 question–answer pairs from 112 of them into an audited pool. Agents turn evaluation requests into specifications, and a deterministic selector with a red-team gate composes reproducible benchmarks. We release Video-Index, the 210 hardest verified items under these attacks in each of four capability groups, 840 items from 76 sources. With the same fixed input, Claude Opus 5 outscores every open-source model by over 37 percentage points, and agent tools add about 20 more, yet all systems leave room to improve efficiency and accuracy.

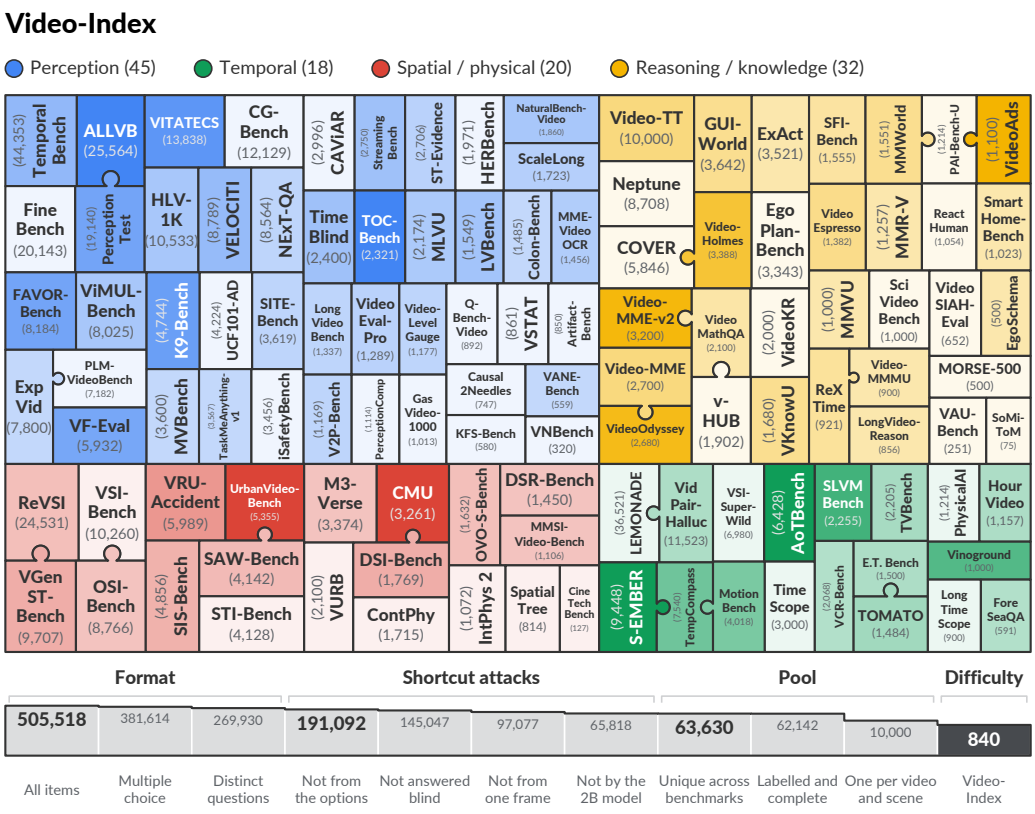


Figure 1: **Screening keeps 840 of 505,518 pooled items.** Top, the 115 benchmarks by capability group, sized by log item count, shaded by Video-Index items. Bottom, items per stage.

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.

1 INTRODUCTION

A benchmark score is a certificate of a capability. A video model is trusted to follow a soccer match, an egocentric kitchen, or an hour-long lecture once it scores well on the matching benchmarks. These test temporal order (Cores et al., 2024; Liu et al., 2024c), long video (Chandrasegaran et al., 2024; Wang et al., 2024a), spatial reasoning (Yang et al., 2025b), egocentric memory (Mangalam et al., 2023), knowledge (Hu et al., 2026), hallucination (Wang et al., 2024b), streaming (Lin et al., 2024), and general perception (Li et al., 2024b; Fu et al., 2025). In the past five years, 605 video benchmarks have appeared, 342 of them with public data. The 115 multiple-choice benchmarks we evaluate hold 528,172 items over 123,281 videos, as many as the instruction-tuning sets behind the first video language models (Maaz et al., 2024; Zhang et al., 2024c). Yet each benchmark was built against the models of its year. Stronger language priors, richer single frames, and retrieval over items already asked now let a model earn the certificate without watching the video.

We argue that a certificate is worth the difficulty of obtaining it without the capability, measured for every benchmark under one protocol. An adversary takes every item it can solve without the capability, and the benchmark keeps the rest. The attack pyramid organizes five attacker sets of growing access. The option level sees the option layout, the text level adds the question, and the pool level adds the other evaluated items. The frame level adds one frame or a caption, and the order level adds a shuffled or truncated video. The exploitability of a benchmark at a level is the largest margin over chance that its attackers reach. The benchmark breaks at the lowest level whose attackers approach the full protocol. Prior audits probe one benchmark with one shortcut (Buch et al., 2022; Cores et al., 2024; Shangguan et al., 2024; Yang et al., 2026a; Brown et al., 2025) and discard it when the shortcut works. Generators write new items with agents (Zhang et al., 2024b; Butt et al., 2024; LI et al., 2025; Shi et al., 2026a) and leave existing benchmarks unused. The pyramid instead tests every benchmark at every level and keeps what no tested attack solves.

We run the pyramid over 115 video benchmarks. Newer benchmarks break earlier, each claim breaks at its own level, and the errors of most benchmarks concentrate in one or two capabilities. We keep the items that no attacker solves and let agents compose benchmarks from them behind a red-team gate. As shown in Figure 1, screening keeps 840 of the 505,518 items in the pool for Video-Index, a meta-benchmark of the hardest items whose answers a judge verified against the frames. We evaluate open-source video models and proprietary agents on Video-Index.

- We introduce the attack pyramid, run it on 115 video benchmarks under one protocol, and find that newer benchmarks break earlier and that each claim breaks at its own level.
- Our agent-assisted pipeline combines screening, specification generation, deterministic selection, and red-team validation for reproducible, on-demand benchmarks with lockfiles.
- We release Video-Index, a meta-benchmark of the 840 hardest verified items of the pool, and evaluate open-source video models and proprietary systems on it.

2 WHAT SCORES CERTIFY

We treat a score as evidence for the capability its benchmark claims to test. An attacker raises the score without that capability. An options-only model exploits answer wording, and a single-frame reader skips temporal order. Such shortcuts (Khandelwal et al., 2019; Buch et al., 2022; Zheng et al., 2024a; Balepur et al., 2024) motivate an adversarial test of what the score certifies.

Let B be a benchmark, c its chance accuracy, and $\mathcal{A}$ a declared set of attackers. Write $\mathrm{acc}_B(a)$ for the accuracy of attacker a on B. The exploitability of B is

$$\varepsilon_{\mathcal{A}}(B) = \max_{a \in \mathcal{A}} \mathrm{acc}_B(a) - c, \tag{1}$$

the largest margin over chance that an attacker in the set achieves. The maximizing attacker can differ across benchmarks. For a model with accuracy s, the residual $s - c - \varepsilon_{\mathcal{A}}(B)$ measures the score above the strongest tested shortcut, and a positive residual rules out only the shortcuts tested.

We nest the attacker sets as $\mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{A}_4$. The option level grants options alone, the text level adds the question, and the pool level adds other evaluation items. The frame level admits a degraded visual view, and the order level admits temporally perturbed video. No attacker sees the ground truth of the item under evaluation. We write ε_ℓ for the exploitability at level ℓ, and the cumulative profile $\varepsilon_{\mathrm{opt}}, \varepsilon_{\mathrm{text}}, \varepsilon_{\mathrm{pool}}, \varepsilon_{\mathrm{frame}}, \varepsilon_{\mathrm{order}}$ is nondecreasing. Following Finzi et al. (2026), each level represents an observer with bounded access and compute, as developed in the appendix.

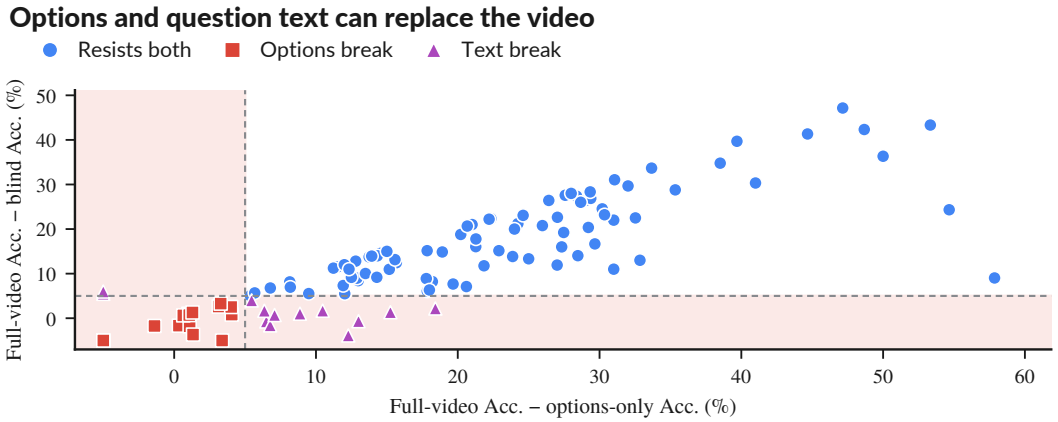


Figure 2: **Options and question text replace the video on 30 benchmarks.** Axes give full-video accuracy minus options-only and minus blind accuracy, and bands mark breaks.

A benchmark's threat model specifies the levels its capability claim must resist. General video understanding (Li et al., 2024b; Liu et al., 2024c; Cai et al., 2024; Fu et al., 2025) requires every level, and a knowledge claim (He et al., 2024; Hu et al., 2026) could exempt the text level, yet we apply every measured level to every benchmark. Let the reference accuracy s^* be the full-protocol accuracy of a reference model outside the attacker sets. A benchmark breaks when $\varepsilon_\ell \geq s^* - c - \delta$, with tolerance $\delta = 5\%$, about the 95% sampling margin of a 300-item accuracy, and its breaking level is the first such level. An unbroken benchmark resists every tested level within this tolerance.

3 The Attack Pyramid

3.1 Benchmarks and Protocol

We census 605 video benchmarks released in the past five years, 342 of them with public data, and collect 721,324 question–answer pairs across 161,580 videos. We retain 115 English multiple-choice benchmarks that do not require audio. We sample $n = 300$ questions per benchmark, preserving subtask proportions and keeping every question when fewer are available. Videos are stored at a short side of 720 pixels, reduced to 480 for clips longer than 30 minutes without on-screen text. Frames reach the models with a long side of at most 768 pixels. Gemini 3.5 Flash (Google DeepMind, 2026b) supplies the reference accuracy s^* at 32 frames. On long-video benchmarks where it stays within the tolerance δ of chance, its 128-frame probe supplies the reference instead. **Models in this audit answer once per item, without tools, retries, or an agent scaffold.** We label items with 18 fine capability categories in four groups, namely perception, temporal, spatial and physical, and reasoning and knowledge. Each benchmark takes the group that most of its items require. Before attacks, we score each answer by its option letter, then by an answer-line regular expression, then by its option text. Within each benchmark, we keep one item per near-duplicate group, which removes 1.1% of sampled items, refill to $n = 300$ where possible, and screen the replacements.

3.2 Options without Questions

We test whether the answer options alone reveal the answer. Chance accuracy is the reciprocal of the option count of each item. A fixed-position attacker picks the most frequent answer position. An options-only attacker, Claude Opus 5 (Anthropic, 2026b), sees only the choices in a call that shares no context with its other roles. The larger margin over chance defines ε_{opt}. As shown in Figure 2, options raise accuracy by up to 63% over chance, lift it significantly on 78 benchmarks, and break 17. Across 70 complete runs, eight permutations of the option order give Qwen3-VL-8B (Bai et al., 2025) a median accuracy range of 5%, significant on 15 benchmarks. The selected answer changes on a median 37% of items and on 82% of DSR-Bench (Zhou et al., 2025d).

3.3 Questions without Video

We test whether the question reveals the answer without the video. Gemini 3.5 Flash (Google DeepMind, 2026b) and Claude Opus 5 (Anthropic, 2026b) read the question and options independently, and their larger margin over chance defines $\varepsilon_{\text{text}}$. Video dependence m is reference accuracy minus the accuracy of the stronger blind reader, and a larger m marks stronger resistance to text shortcuts. As shown in Figure 2, question text raises accuracy by up to 71% over chance, against 63% for options alone. Among the 98 benchmarks that pass the option level, video dependence spans 51%. The blind reader exceeds reference accuracy on 4 of them and breaks 13.

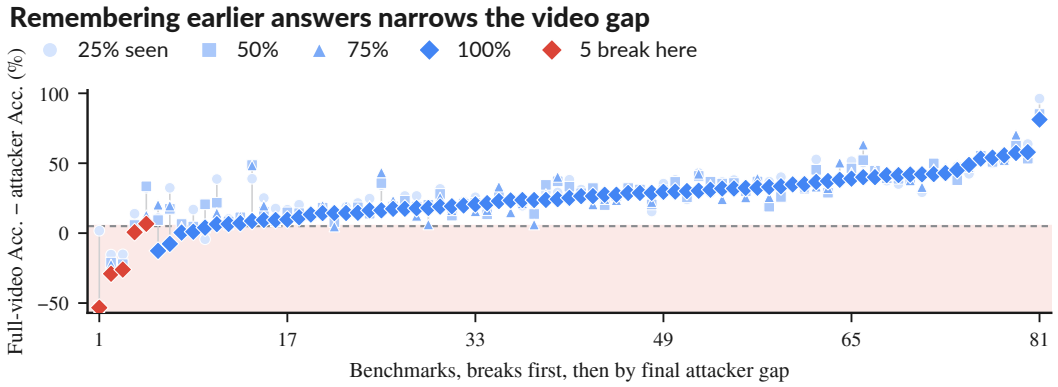


Figure 3: **Remembering earlier answers narrows the video gap.** Each column is a benchmark, with its gap to the strongest pool attacker after 25 to 100% of the items.

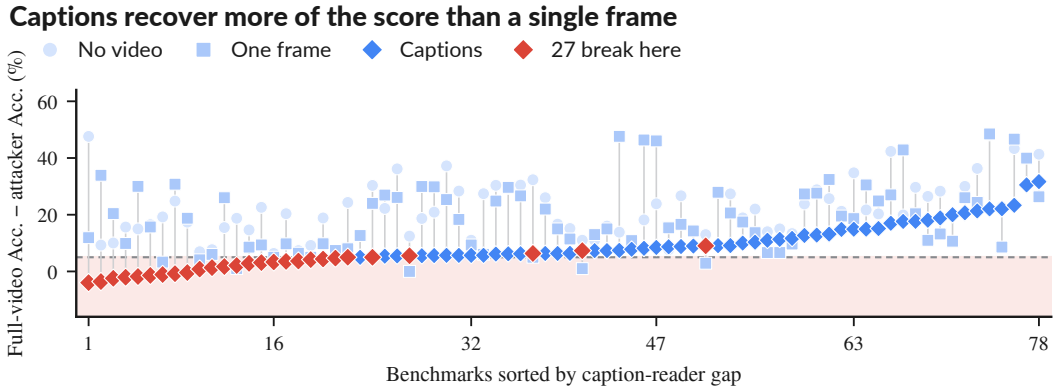


Figure 4: **Captions recover more of the video score than a single frame.** Each column is a benchmark, sorted by its caption gap, with gaps under no video, one frame, and captions.

3.4 Learning from evaluation items

We test whether the answers to other items help solve a question. A copy attacker copies the answer of the nearest earlier question when their BGE-large (Xiao et al., 2024) cosine similarity exceeds 0.9. A retrieval attacker gives Qwen3-8B (Yang et al., 2025a) the eight nearest question–answer pairs. Following the predictive auditing of Brown et al. (2025), an online logistic regression on BGE-large embeddings predicts answer letters out of sample over five rotating splits. As shown in Figure 3, the gap to the strongest attacker narrows as more items enter the sequence, and 5 benchmarks break.

3.5 Single frames and captions

We replace the video with a single frame or a text description. Gemini 3.5 Flash (Google DeepMind, 2026b) reads one frame, and Claude Opus 5 (Anthropic, 2026b) reads captions of the 32 sampled frames, each written independently by AuroraCap-7B-IMG (Chai et al., 2025) without the question or options. The larger margin over chance defines $\varepsilon_{\text{frame}}$. As shown in Figure 4, captions recover more reference accuracy than a single frame. Among the 78 benchmarks with all three probes at this level, one frame comes within the tolerance δ of the reference on 8. It beats the blind reader by a median 3.3%. Captions come within δ on 23 and exceed the reference on 9, and 27 benchmarks break at this level.

3.6 Order and temporal coverage

We test whether accuracy depends on frame order and on coverage of the full clip. Gemini 3.5 Flash (Google DeepMind, 2026b) and Qwen3-VL-8B (Bai et al., 2025) read 32 shuffled frames and, separately, a contiguous tenth of the video, and their largest margin over chance defines $\varepsilon_{\text{order}}$. As shown in Figure 5, much of the score survives both perturbations. Among the 51 benchmarks with both probes at this level, shuffling retains a median 96% of the same model's full-video accuracy and one tenth retains 78%. Across all 96 profiled benchmarks, either half retains a median 92%, and shifting the sampling grid leaves the median unchanged. Of 15 benchmarks whose papers claim temporal reasoning, 6 retain over 90% under shuffling, yet 6 of the 7 reaching this level lose at least a quarter. In total, 38 benchmarks remain unbroken.

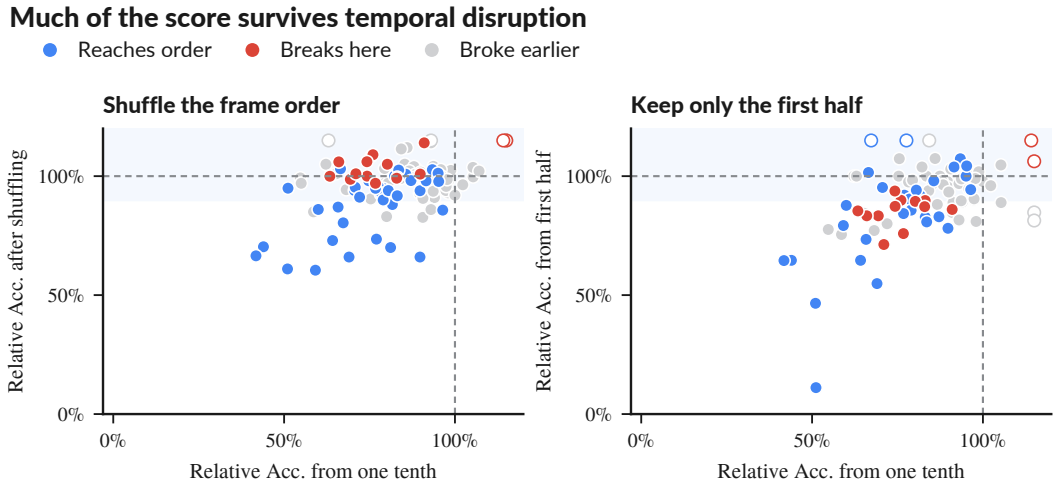


Figure 5: **Much of the score survives temporal perturbation.** Each dot is a benchmark, placed by the share of full-video accuracy that each perturbation retains.

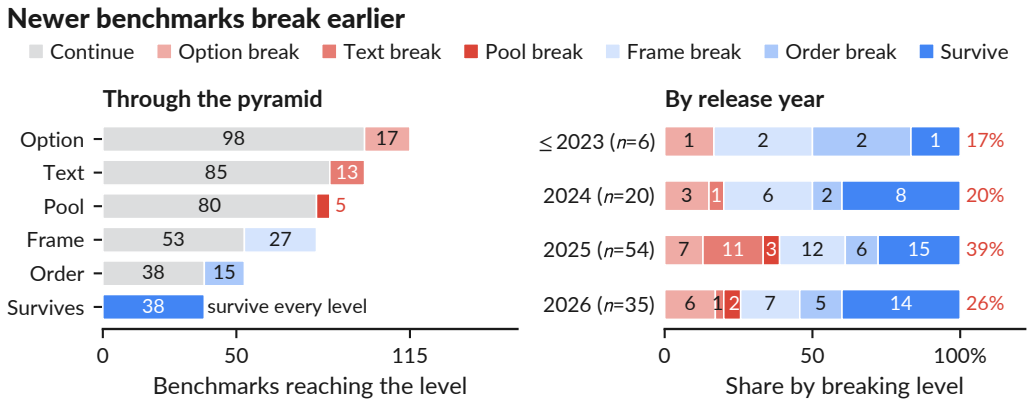


Figure 6: **Newer benchmarks break earlier.** Left, benchmarks that reach and break at each level. Right, breaking levels by release year, with the share breaking before visual input in red.

4 What the audit finds

4.1 Newer benchmarks break earlier

We group the breaking levels by release year. As shown in Figure 6, newer benchmarks are easier to attack without video. Of the 115 audited benchmarks, 38 survive every level and 35 break before any visual input. The share breaking before visual input rises from 19% among releases through 2024 to 34% among releases in 2025 and 2026. The rise comes from the text and pool levels, where releases from 2025 and 2026 hold 17 of the 18 breaks. Blind readers break 12 of the 89 newer benchmarks against 1 of the 26 older ones. On most of them the blind reader comes within the tolerance δ of a reference that scores far above chance. The option level instead breaks a similar share in both periods, 4 older and 13 newer benchmarks, especially where reference accuracy approaches chance. The rise persists under attacks matched to each claim, as the appendix details. Recent benchmarks use LLMs to generate or refine questions and answers (Liu et al., 2024c; Hu et al., 2026), which can make the answer recoverable from text alone. This may also leave recurring patterns in answer options, creating shortcuts even when questions resist blind reading.

4.2 Each claim breaks differently

As shown in Figure 7, the four capability groups break at different levels. Reasoning and knowledge benchmarks break on text or captions that already carry the facts. Perception benchmarks break most often at the frame level, and spatial and physical ones on options and under temporal perturbation. Temporal benchmarks split sharply, with 6 of 18 breaking on options and 10 surviving every level. Audits should match their checks to the claimed capability and still screen options for temporal claims. The appendix provides the finer breakdown within each broad capability group.

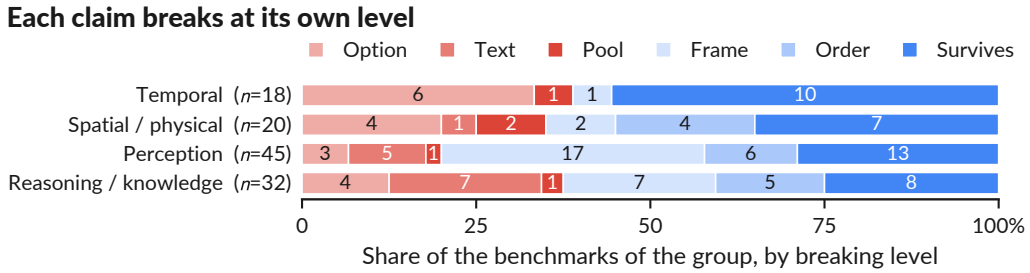


Figure 7: **Each capability group breaks at its own levels.** Bars split the benchmarks of each group by breaking level, and segment labels give counts.

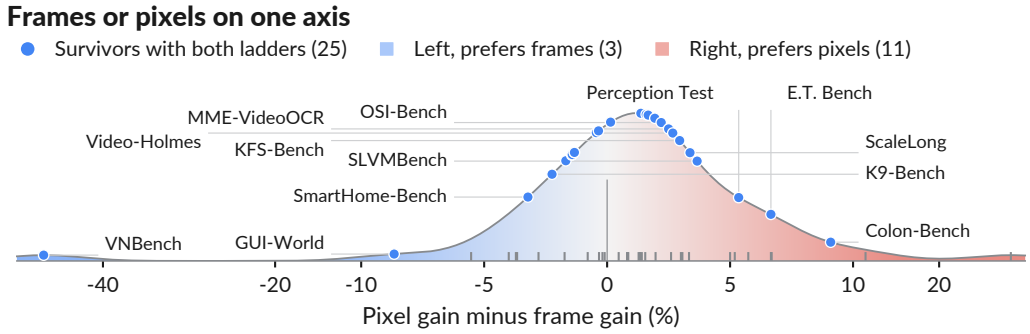


Figure 8: **Eleven benchmarks gain more from pixels and three from frames.** The curve is the density of pixel gain minus frame gain, and dots mark survivors with both ladders.

4.3 Frames or pixels

Additional visual input can supply finer spatial detail or denser temporal coverage, and we ask which helps more. On the benchmarks that pass the pool level, Qwen3-VL-8B (Bai et al., 2025) reads 32 frames at short sides of 168, 224, 336, and 448 pixels and at the stored resolution. It then reads 224-pixel frames sampled at 0.25 to 2 frames per second within its context. As shown in Figure 8, 11 benchmarks favor pixels, 3 favor frames, and the two gains differ by less than δ on the rest. Resolution recovers small details in slides, endoscopy, and synthetic artifacts, and denser sampling catches brief events that sparse sampling misses.

We also test whether long-video questions need evidence that sparse sampling misses, with frame budgets from 32 to 1024. As shown in Figure 9, most long-video benchmarks gain from larger budgets. Among the 30 benchmarks whose videos average at least 300 seconds, 18 gain at least δ in the reference sweep and 9 still gain at its 600-frame limit. These gains indicate that sparse inputs can omit evidence useful for answering. In a diagnostic sweep at 224 pixels, only LongTimeScope (Zohar et al., 2025a) and TimeScope (Zohar et al., 2025b) keep improving from 512 to 1024 frames. Since the observed gains also depend on the reference model's long-context ability, other benchmarks may still place high demands on processing events across long time spans.

4.4 Errors cluster in few capabilities

We attribute the 32-frame errors of the reference model to the 18 fine categories of Section 3.1, on the 76 pool-level survivors with at least 20 agreed attributions. Claude Opus 5 (Anthropic, 2026b) receives the 32 frames, question, options, gold answer, and incorrect answer with its reasoning, and only attributions confirmed by two independent judgments count. As shown in Figure 10, errors concentrate in few capabilities. The two most frequent categories of a benchmark hold a median 71% of its attributions, and profiles overlap weakly, with mean cosine similarity 0.29. Fine-grained action leads on 27 benchmarks, spatial relation on 13, and temporal order and action counting on 7 each, counting ties. Covering the observed capability gaps therefore takes several sources.

4.5 Duplicates reduce effective size

We count distinct items, since item count can overstate diversity (Friedman & Dieng, 2022). Let $\mathbf{v}_i$ be the unit-normalized mean SigLIP2 (Tschannen et al., 2025) embedding of the 32 video frames of item i, and let $\mathbf{t}_i$ be its unit-normalized BGE-large (Xiao et al., 2024) question embedding. For a sample of n items, the joint similarity matrix K and its Vendi score are

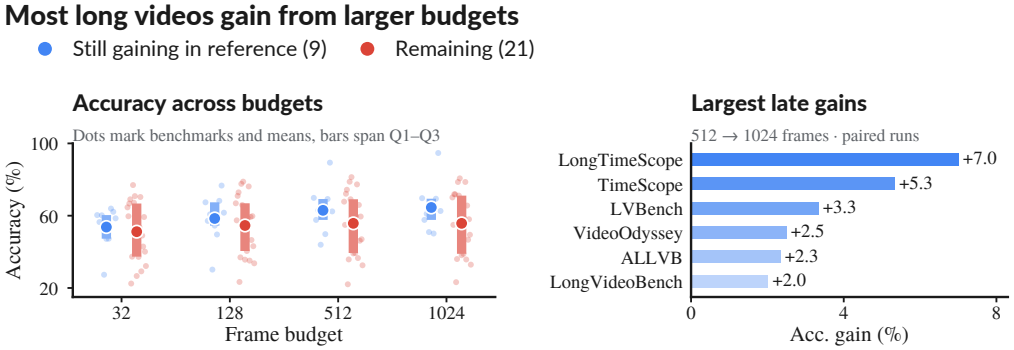


Figure 9: **Most long videos gain from larger budgets.** Left, scores per budget with means and quartiles, blue if the reference still gains at 600 frames. Right, the six largest gains, 512 to 1024.

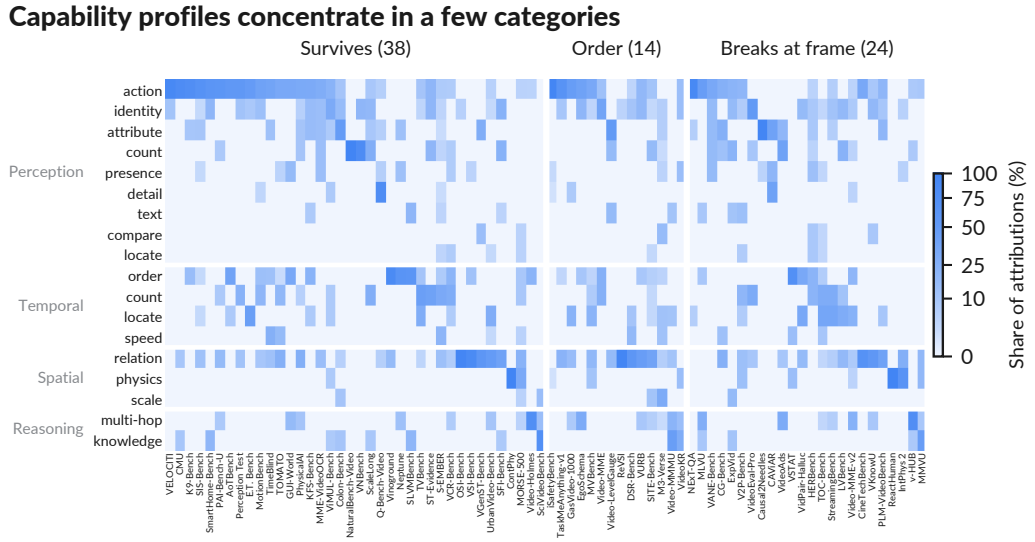


Figure 10: **Errors concentrate in few capabilities.** Columns are the 76 pool-level survivors with 20 or more agreed attributions, grouped by their first breaking level, each summing to 100%.

$$K_{ij} = (\mathbf{v}_i^\top \mathbf{v}_j)(\mathbf{t}_i^\top \mathbf{t}_j), \qquad \mathrm{VS}(K) = \exp\Big(-\sum_{k=1}^{n} \lambda_k \log \lambda_k\Big), \tag{2}$$

where λ_k is the kth eigenvalue of K/n. The Vendi score counts how many distinct items carry the observed diversity, and with $n = 200$ we call it the effective size n_{eff}. As shown in Figure 11, a 200-item draw holds a median of 32 effective items, and 39 benchmarks hold 20 or fewer. Near-duplicate questions make up at least half the items of 63 benchmarks, and weighting items by inverse duplicate-cluster size moves accuracy by more than 1% on 93 benchmarks. Across sources, a tenth of items use videos that appear in another benchmark. PAI-Bench-U (Zhou et al., 2026) and PhysicalAI (Zhou et al., 2026) share 191 questions and yield nearly identical scores across conditions.

5 Rebuilding Video Evaluation

5.1 Screening Every Item

We screen all 381,614 multiple-choice items from 112 of the 115 audited benchmarks, leaving out the three added most recently. Qwen3-VL-8B (Bai et al., 2025) removes items it answers from text alone, options alone, or a single frame, and Qwen3-VL-2B (Bai et al., 2025) removes items it answers from 32 frames in order or shuffled. We then drop near-duplicate questions across sources and restrict repeated videos during selection. A labeling agent assigns task and video-content labels, and we record provenance, duration, attack margins, and question and middle-frame embeddings. As shown in Figure 1, attacks remove 76% of distinct questions, and 62,142 remain with complete screening and labeled videos. From these, the selector draws a 10,000-item coverage composition over 105 benchmarks, four capability groups, and 19 content categories.

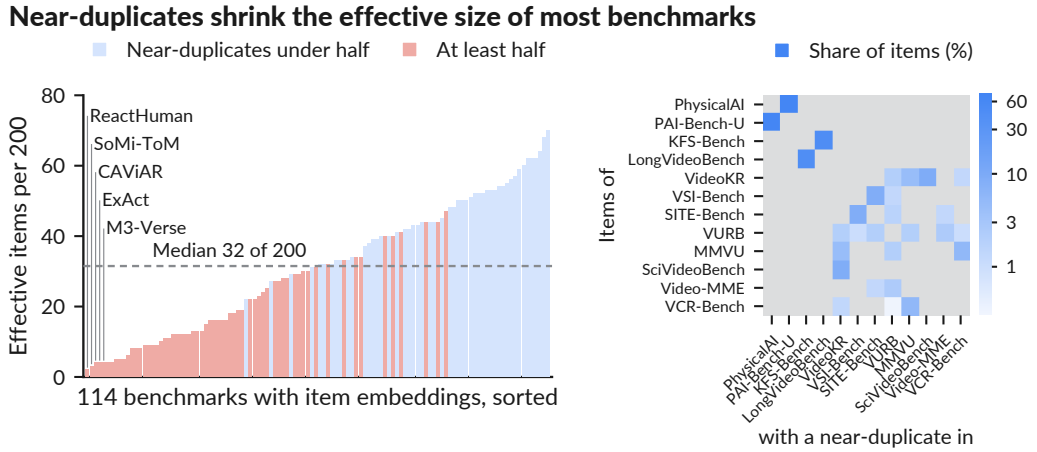


Figure 11: **Near-duplicates shrink effective size.** Left, effective items per 200, red where at least half are near-duplicates. Right, the largest near-duplicate shares across benchmarks.

Table 1: **Agents lead open models on Video-Index.** Accuracy in %, fixed input then agent tools. Blind sees text only and Gain is Video minus Blind. Highest in bold. Protocols in the appendix.

Model	Video	Blind	Gain	Percep.	Temp.	Spatial	Reason.
VideoChat3-4B	9.4	3.8	5.7	9.5	12.4	6.2	9.5
Qwen3-VL-8B	9.5	1.4	8.1	9.0	11.4	6.7	11.0
Cambrian-S-7B	11.3	5.3	6.0	12.4	10.0	3.8	19.0
InternVideo2.5-8B	12.4	7.5	4.9	12.9	11.0	8.1	17.6
VideoLLaMA3-7B	12.4	5.2	7.1	11.9	18.6	8.6	10.5
Video-XL-2	12.4	7.3	5.1	12.4	15.7	9.0	12.4
LLaVA-OneVision-2-8B	14.2	1.8	12.4	11.4	21.0	11.9	12.4
InternVL3.5-8B	15.2	4.4	10.8	13.3	20.0	8.1	19.5
Molmo2-8B	19.0	4.8	14.2	20.0	16.2	17.1	22.9
Claude Opus 5	56.8	21.3	35.5	62.4	56.2	46.7	61.9
Agentic evaluation							
Gemini 3.5 Flash-Lite	22.5	17.5	5.0	20.0	30.0	10.0	30.0
Gemini 3.1 Pro	54.7	23.9	30.8	58.2	59.1	60.6	40.9
Claude Opus 5	70.6	**24.9**	45.7	68.4	77.2	69.1	67.6
Claude Fable 5.1	71.5	20.7	50.8	68.4	77.7	71.1	68.7
GPT-6-Astra	**79.3**	18.7	**60.6**	**75.4**	**83.7**	**80.4**	**77.1**
Human	55.0	17.5	37.5	56.0	68.0	56.0	40.0

5.2 Composition from a request

A benchmark request specifies task and event coverage, an item budget k, and a cap per source. Following agent-assisted benchmark construction (Butt et al., 2024; LI et al., 2025), a specification agent translates the request into a query over the pool and returns it for confirmation. A request can, for example, select household videos longer than 60 seconds with planning or memory questions. For the coverage composition, the selector draws items round-robin across duration groups, event types, and sources. Its initial caps allow one item per video and visual scene, five per question template, and 5% of the budget per source, and scarce coverage cells trigger recorded relaxations. Within a source, the selector prefers less templated questions and then weaker worst-case attacker scores. The composition holds 9,688 videos from 105 benchmarks across seven duration groups. The appendix details this composition and its evaluation.

After selection, a red-team gate applies adversarial filtering (Le Bras et al., 2020) to check the composition for residual shortcuts. It tests selected items with Qwen3-VL-8B (Bai et al., 2025) at 32 frames and the caption attacker, then reruns all five levels. It drops solved items and refills for a bounded number of rounds. A registrar randomizes option order, reports residual exploitability, and logs the pool snapshot, query, selector version, and seed in a lockfile whose hash names the release.

5.3 Video-Index

Video-Index stress-tests models on the 840 hardest verified questions from 76 sources, 210 per capability group and one per video, ranked by their maximum percentile over five attacker margins. Claude Opus 5 (Anthropic, 2026b) labels the 3,000 hardest candidates by fine capability and verifies their answers against the frames, and selection ignores reference-model accuracy. We evaluate nine open video models, namely Qwen3-VL-8B (Bai et al., 2025), VideoChat3-4B (Li et al., 2026d), Cambrian-S-7B (Yang et al., 2026a), InternVideo2.5-8B (Wang et al., 2025d), VideoLLaMA3-7B

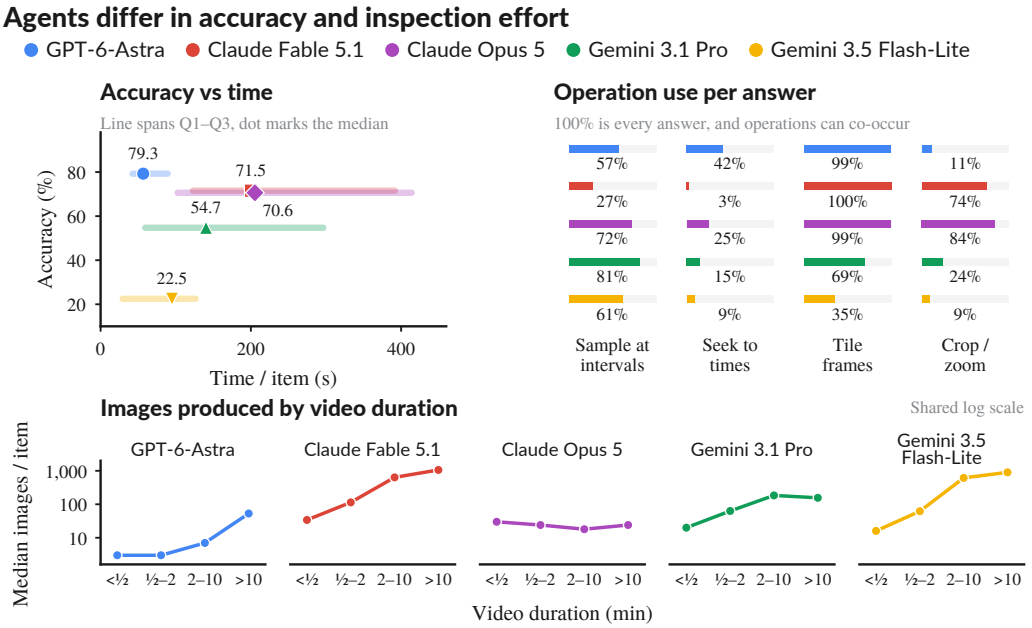


Figure 12: **Agents differ in accuracy, runtime, and inspection.** Left, accuracy and median runtime, Q1–Q3. Right, operations per answer. Bottom, image medians by duration on shared log axes.

(Zhang et al., 2025a), Video-XL-2 (Qin et al., 2025), LLaVA-OneVision-2-8B (An et al., 2026), InternVL3.5-8B (Wang et al., 2025a), and Molmo2-8B (Clark et al., 2026). These models and a fixed-input Claude Opus 5 receive at most 512 frames at 1 fps with a 224-pixel short side. Five agents use local tools on the same 2-fps videos, namely GPT-6-Astra (OpenAI, 2026), Claude Fable 5.1 (Anthropic, 2026a), Claude Opus 5 (Anthropic, 2026b), Gemini 3.1 Pro (Google DeepMind, 2026a), and Gemini 3.5 Flash-Lite (Google DeepMind, 2026c). Blind scores average four option orders. As shown in Table 1, fixed-input Claude Opus 5 outscores every open model by more than 37%. On matched items, tools raise the same model by 19.8%. Human volunteers reach 55.0%, below the three leading agents and above every open model. At a matched item budget, the gate holds the strongest attacker to 13%, against 19% for ungated selections. Blind baselines (Chen et al., 2024b) give a mean text-only accuracy of 6% and a video gain of 11% over ten models, with video above blind for all ten. Item-bootstrap signal-to-noise analysis (Heineman et al., 2026) gives a noise floor of 2.0% and a signal-to-noise ratio of 10.8, against 3.7 for the median source. The appendix gives protocols, validation, and coverage-first compositions.

We analyze execution logs to see how agents gather evidence. As shown in Figure 12, GPT-6-Astra (OpenAI, 2026), Claude Fable 5.1 (Anthropic, 2026a), and Claude Opus 5 (Anthropic, 2026b) all tile frames into grids on nearly every answer and then diverge. Any two of them give the same answer on about four in five items but choose the same sampling strategy on fewer than half. GPT-6-Astra seeks timestamps more often, while Claude Fable 5.1 and Claude Opus 5 favor cropping and zooming. Claude Opus 5 extracts similar image counts across durations. GPT-6-Astra decodes all stored frames on short clips and combines methods on longer videos, whereas Claude Fable 5.1 decodes exhaustively and extracts far more images without reaching GPT-6-Astra's accuracy. All three score lower on videos past ten minutes, GPT-6-Astra least of all. Evaluation should separate the volume of extracted imagery from its relevance and reward knowing when to stop.

6 Conclusion and Discussion

The attack pyramid exposes shortcuts across 115 video benchmarks and yields Video-Index, 840 of the hardest verified questions. A score certifies a capability only above its strongest shortcut, and builders can report the breaking level beside the score. The screened pool supports further benchmarks with other capability and coverage goals through the same pipeline. Leading proprietary agents outperform open models but leave room to improve accuracy and evidence selection. Resistance holds only against tested attacks, and audits and benchmarks should evolve with new models and shortcuts. Progress requires accuracy, shortcut resistance, and efficient evidence gathering.

AI USE STATEMENT

We used Codex to assist with manuscript drafting and revision, English–Chinese translation, literature and citation checks, figure and table preparation, and editing plotting and LaTeX code. AI-assisted revisions were checked against the reported results, and the manuscript was compiled and visually inspected. Language and vision–language models also serve as experimental components for shortcut attacks, captioning, answer verification, and agent-assisted benchmark construction, as described in the methods and appendices. The authors take responsibility for the final text, code, figures, results, and scientific claims.

Ethics statement. We audit and recombine public benchmarks, and Video-Index reuses items from 76 of them. Each item keeps its source benchmark and provenance, the pool records the license of every source, and a composition can exclude sources by license. The audited videos come from public releases and show people, and some sources contain accidents or crimes. This work records no new footage, and the frames in the appendix come from these releases. Volunteers answered Video-Index items for the human baseline through a browser tool, and the paper reports only their aggregate accuracy. The shortcut analyses show how scores on the audited benchmarks can rise without the claimed capability. The report cards document these shortcuts for each benchmark, and the screened pool removes the items that the tested attackers solve.

Reproducibility statement. Every reported accuracy is computed from item-level records containing the item identifier, sampled frames, model output, and scored answer. The standard audit uses one model call per item. The agentic evaluations additionally record their tool interactions. The release includes sampling seeds, prompts, and snapshot identifiers for each benchmark, condition, and attacker. Figure scripts render the exported results without rerunning experiments. The appendix collects the settings of every stage in one table.

REFERENCES

Raiyaan Abdullah, Yogesh Singh Rawat, and Shruti Vyas. iSafetyBench: A video-language benchmark for safety in industrial environment. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

Raiyaan Abdullah, Shehreen Azad, and Yogesh Singh Rawat. Learning to deny: Action denial in multimodal large language models. In *European Conference on Computer Vision*, 2026.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980. IEEE, 2018.

Reem Aleithan, Haoran Xue, Mohammad Mahdi Mohajer, Elijah Nnorom, Gias Uddin, and Song Wang. Swe-bench+: Enhanced coding benchmark for llms. *arXiv preprint arXiv:2410.06992*, 2024.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Anthropic. Claude Fable 5.1. `https://www.anthropic.com/claude/fable`, 2026a.

Anthropic. Introducing Claude Opus 5. `https://www.anthropic.com/news/claude-opus-5`, 2026b.

Khush Attarde, Yusuf Ali, Megha Thukral, Divye Bhutani, Thomas Ploetz, and Zsolt Kira. K9-Bench: Evaluating multimodal LLMs on canine-centric videos. *arXiv preprint arXiv:2607.02680*, 2026.

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

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do llms answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, 2024.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Himu: Hierarchical multimodal frame selection for long video question answering. *arXiv preprint arXiv:2603.18558*, 2026a.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A benchmark for multi-evidence integration in video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Rohit Bharadwaj, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. VANE-Bench: Video anomaly evaluation benchmark for conversational LMMs. In *Findings of the Association for Computational Linguistics: NAACL*, 2025.

Andy Bonnetto, Haozhe Qi, Franklin Leong, Matea Tashkovska, Mahdi Rad, et al. EPFL-Smart-Kitchen: An ego-exo multi-modal dataset for challenging action and motion understanding in video-language models. In *Advances in Neural Information Processing Systems*, 2025.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should" train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2907–2917. IEEE, 2022.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425, 2024.

Natasha Butt, Varun Chandrasekaran, Neel Joshi, Besmira Nushi, and Vidhisha Balachandran. Benchagents: Multi-agent systems for structured benchmark creation. *arXiv preprint arXiv:2410.22584*, 2024.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Zikui Cai, Andrew Wang, Anirudh Satheesh, Ankit Nakhawa, Hyunwoo Jae, et al. MORSE-500: A programmatically controllable video benchmark to stress-test multimodal reasoning. *arXiv preprint arXiv:2506.05523*, 2025.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 16845–16889, 2025.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, et al. HourVideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, 2024.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, et al. GUI-World: A video benchmark and dataset for multimodal GUI-oriented understanding. In *International Conference on Learning Representations*, 2025a.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, 2025b.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. ReXTime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Weitao Chen, Hu Jiaxin, Xie Tianyidan, Yang Li, Yuyi Qian, Banghao Xu, Ziheng Tang, Shenyi Wang, Mingyue Yu, Duo Li, et al. Egomonth: A month-level egocentric video benchmark for long-term spatiotemporal memory. *arXiv preprint arXiv:2608.13113*, 2026.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, et al. EgoPlan-Bench: Benchmarking multimodal large language models for human-level planning. *International Journal of Computer Vision*, 2025c.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, et al. Scaling RL to long videos. In *Advances in Neural Information Processing Systems*, 2025d.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like Holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, et al. PerceptionLM: Open-access data and models for detailed visual understanding. In *NeurIPS*, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for VideoLLMs. *arXiv preprint arXiv:2410.07752*, 2024.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. SciVideoBench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. SoMi-ToM: Evaluating multi-perspective theory of mind in embodied social interactions. In *Advances in Neural Information Processing Systems*, 2025.

Gueter Josmy Faure, Min-Hung Chen, Jia-Fong Yeh, Hung-Ting Su, and Winston H. Hsu. FineBench: Benchmarking and enhancing vision-language models for fine-grained human activity understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*, 2025.

Marc Finzi, Shikai Qiu, Yiding Jiang, Pavel Izmailov, J Zico Kolter, and Andrew Gordon Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv preprint arXiv:2601.03220*, 2026.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, et al. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026a.

Lin Fu, Zheyuan Yang, Yang Wang, Tingyu Song, Arman Cohan, and Yilun Zhao. VideoKR: Towards knowledge- and reasoning-intensive video understanding. In *International Conference on Machine Learning*, 2026b.

Sparsh Garg, Yi-Wen Chen, Vijay Kumar B G, and Abhishek Aich. CAViAR: A causal video dataset for fine-grained accident reasoning in real-world scenarios. In *European Conference on Computer Vision Workshops*, 2026.

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

Google DeepMind. Gemini 3.5 Flash model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, 2026b.

Google DeepMind. Gemini 3.5 Flash-Lite model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/`, 2026c.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Tianjun Gu, Tianyu Xin, Kuan Zhang, Bowen Yang, Kok-Chung Chua, et al. Towards spatial supersensing in the wild. In *European Conference on Computer Vision*, 2026.

Zhentao Guo, Chen Duan, Tongkun Guan, Zining Wang, Kai Zhou, and Pengfei Yan. Vitexqa: A multi-frame temporal perception dataset for video text question answering. *arXiv preprint arXiv:2606.24602*, 2026.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018.

Abdullah Hamdi, Changchun Yang, and Xin Gao. Colon-Bench: An agentic workflow for scalable dense lesion annotation in full-procedure colonoscopy videos. In *Medical Image Computing and Computer Assisted Intervention*, 2026.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, et al. VideoEspresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Haichen He, Jiayi Zhou, Sifeng Shang, Yihan Hu, Yuanhan Zhang, and Kaiyang Zhou. VideoOdyssey: A benchmark for ultra-long-context and omni-modal video understanding. *arXiv preprint arXiv:2605.22907*, 2026.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah Smith, Hanna Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. *Advances in Neural Information Processing Systems*, 38:17073–17114, 2026.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, et al. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026.

Haojian Huang, Harold Haodong Chen, Meng Luo, Junjia Du, Shanqing Xu, et al. No place to hide: Benchmarking video hallucination with background-controlled pairs. In *European Conference on Computer Vision*, 2026a.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. Eg-vqa: Benchmarking verifiable video question answering with grounded temporal evidence. *arXiv preprint arXiv:2606.24797*, 2026b.

Tianxiang Jiang, Sheng Xia, Yicheng Xu, Linquan Wu, Xiangyu Zeng, Limin Wang, Yu Qiao, and Yi Wang. VKnowU: Evaluating visual knowledge understanding in multimodal LLMs. *arXiv preprint arXiv:2511.20272*, 2025.

Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, et al. Datbench: Discriminative, faithful, and efficient vlm evaluations. *arXiv preprint arXiv:2601.02316*, 2026.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124, 2021.

Younggun Kim, Ahmed S. Abdelrahman, and Mohamed Abdel-Aty. VRU-Accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2025.

Alex Kipnis, Konstantinos Voudouris, Luca Schulze Buschoff, and Eric Schulz. metabench-a sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 31734–31770, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*, pp. 1078–1088. Pmlr, 2020.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, 2023.

Baiqi Li. NaturalBench-Video (dataset). `https://huggingface.co/datasets/BaiqiL/NaturalBench-Video`, 2025.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean De Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024a.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. TimeBlind: A spatio-temporal compositionality benchmark for video LLMs. *arXiv preprint arXiv:2602.00288*, 2026a.

Chuhan Li, Rilyn Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. SAW-Bench: Learning situated awareness in the real world. In *International Conference on Machine Learning*, 2026b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024b.

Miaoyu Li, Qin Chao, and Boyang Li. Two causally related needles in a video haystack. In *Advances in Neural Information Processing Systems*, 2025.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, 2024c.

Shaoxuan Li, Zhixuan Zhao, Hanze Deng, Zirun Ma, Shulin Tian, et al. PerceptionComp: A video benchmark for complex perception-centric reasoning. *arXiv preprint arXiv:2603.26653*, 2026c.

Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. VITATECS: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, 2024d.

XIANG LI, Farzaan Kaiyom, Evan Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. Autobencher: Towards declarative benchmark construction. In *International Conference on Learning Representations*, volume 2025, pp. 94357–94386, 2025.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026d.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. OVO-S-Bench: A hierarchical benchmark for streaming spatial intelligence in multimodal LLMs. In *Conference on Empirical Methods in Natural Language Processing*, 2026e.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? In *International Conference on Computer Vision*, 2025.

Zongyao Li, Kengo Ishida, Satoshi Yamazaki, Xiaotong Ji, and Jianquan Liu. KFS-Bench: Comprehensive evaluation of key frame sampling in long video understanding. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026f.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, et al. MMSI-Video-Bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025.

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

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024a.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. E.T. Bench: Towards open-ended event-level video-language understanding. In *Advances in Neural Information Processing Systems*, 2024b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024c.

David Ma, Huaqing Yuan, Xingjian Wang, Qianbo Zang, Tianci Liu, et al. ScaleLong: A multi-timescale benchmark for long video understanding. *arXiv preprint arXiv:2505.23922*, 2025.

Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. VideoEval-Pro: Robust and realistic long video understanding evaluation. *Transactions on Machine Learning Research*, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

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

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*, 2024.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

OpenAI. GPT-6 Astra: A new generation of intelligence. `https://openai.com/index/gpt-6-astra/`, 2026.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8376–8405, 2026.

Hyojin Park, Yi Li, Janghoon Cho, Sungha Choi, Jungsoo Lee, et al. ForeSea: AI forensic search with multi-modal queries for video surveillance. In *European Conference on Computer Vision*, 2026a.

Jinho Park, Youbin Kim, Hogun Park, and Eunbyung Park. VGenST-Bench: A benchmark for spatio-temporal reasoning via active video synthesis. *arXiv preprint arXiv:2605.22570*, 2026b.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, et al. Perception Test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Benchmark agreement testing done right: A guide for llm benchmark evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2024.

Jason Phang, Angelica Chen, William Huang, and Samuel R Bowman. Adversarially constructed evaluation sets are more challenging, but may not be fair. *arXiv preprint arXiv:2111.08181*, 2021.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pp. 180–191, 2018.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Ameya Prabhu, Vishaal Udandarao, Philip H Torr, Matthias Bethge, Adel Bibi, and Samuel Albanie. Efficient lifelong model evaluation in an era of rapid progress. *Advances in Neural Information Processing Systems*, 37:74089–74121, 2024.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, et al. VCR-Bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. *arXiv preprint arXiv:2506.19225*, 2025.

Qwen Team. Qwen2.5: A party of foundation models. `https://qwenlm.github.io/blog/qwen2.5/`, September 2024.

Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. VideoMathQA: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.

ReactHuman Benchmark Team. ReactHuman benchmark (scaled) (dataset). `https://huggingface.co/datasets/Alan123/reacthuman-benchmark-scaled-with-videos`, 2026.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. VELOCITI: Benchmarking video-language compositional reasoning with strict entailment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, Muhammad Jehanzeb Mirza, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, et al. Livexiv-a multimodal live benchmark based on arxiv papers content. In *International Conference on Learning Representations*, volume 2025, pp. 11470–11502, 2025.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, et al. A culturally-diverse multilingual multimodal video benchmark & model. *arXiv preprint arXiv:2506.07032*, 2025.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*, 2024.

Lin Shi, Haowei Lin, Zixuan Zhu, Xiaoyue Zhou, Xiang Li, Xiangning Lin, Yaxuan Deng, Han Xu, Yuangang Li, Shanda Li, et al. Harbor adapters and harbor-index: Infrastructure and a curated meta-dataset for large-scale agentic evaluation. *arXiv preprint arXiv:2609.04298*, 2026a.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, et al. MME-VideoOCR: Evaluating OCR-based capabilities of multimodal LLMs in video scenarios. In *Advances in Neural Information Processing Systems*, 2025.

Zhengpeng Shi, Yanpeng Zhao, Jianqun Zhou, Yuxuan Wang, Qinrong Cui, et al. v-HUB: A benchmark for video humor understanding from vision and sound. In *Annual Meeting of the Association for Computational Linguistics*, 2026b.

Michal Shlapentokh-Rothman, Prachi Garg, Yu-Xiong Wang, and Derek Hoiem. Decomposing queries into tool calls for long-video keyframe retrieval. *arXiv preprint arXiv:2605.23826*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Tingyu Song, Tongyan Hu, Guo Gan, and Yilun Zhao. VF-Eval: Evaluating multimodal LLMs for generating feedback on AIGC videos. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. ALLVB: All-in-one long video understanding benchmark. In *AAAI Conference on Artificial Intelligence*, 2025.

Yuwen Tan, Joey Huang, Jin Huang, Haoxiang Li, and Boqing Gong. Natural language camera movement understanding. In *European Conference on Computer Vision*, 2026.

Yuqi Tang, Yang Shi, Zhuoran Zhang, Qixun Wang, Xuehai Bai, et al. Artifact-Bench: Evaluating MLLMs on detecting and assessing the artifacts of AI-generated videos. *arXiv preprint arXiv:2605.18984*, 2026a.

Ziyao Tang, Pengkun Jiao, Bin Zhu, Huiyan Qi, Jingjing Chen, and Yu-Gang Jiang. Spatiotemporal sycophancy: Negation-based gaslighting in video large language models. *arXiv preprint arXiv:2604.17873*, 2026b.

TOC-Bench Team. TOC-Bench: A temporal object consistency benchmark for video large language models (dataset). `https://huggingface.co/datasets/anonymous-video-benchmark/toc_bench`, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. FAVOR-Bench: A comprehensive benchmark for fine-grained video motion understanding. *arXiv preprint arXiv:2503.14935*, 2025.

Hao Wang, Hanchen Li, Qiuyang Mang, Alvin Cheung, Koushik Sen, and Dawn Song. Do androids dream of breaking the game? systematically auditing ai agent benchmarks with benchjack. *arXiv preprint arXiv:2605.12673*, 2026a.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-backed video question answering. In *European Conference on Computer Vision*, 2026b.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, et al. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024a.

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. *arXiv preprint arXiv:2508.18265*, 2025a.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, et al. SITE: towards spatial intelligence thorough evaluation. In *International Conference on Computer Vision*, 2025b.

Xiaodong Wang, Xuanyi Zhao, Pedro Rodriguez, Devendra Singh Sachan, Barlas Oguz, et al. S-EMBER: A large-scale benchmark for streaming egocentric memory retrieval. *arXiv preprint arXiv:2607.02689*, 2026c.

Xinran Wang, Songyu Xu, Xiangxuan Shan, Yuxuan Zhang, Muxi Diao, et al. CineTechBench: A benchmark for cinematographic technique understanding and generation. *arXiv preprint arXiv:2505.15145*, 2025c.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025d.

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User centric continual spatial intelligence reasoning in egocentric video streams. *arXiv preprint arXiv:2606.15200*, 2026d.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024b.

Kewei Wei, Bocheng Hu, Jie Cao, Xiaohan Chen, Zhengxi Lu, et al. M^3-Verse: A "spot the difference" challenge for large multimodal models. *arXiv preprint arXiv:2512.18735*, 2025.

Yuancheng Wei, Linli Yao, Lei Li, Haojie Zhang, Hao Zhou, Fandong Meng, and Xu Sun. Video understanding reward modeling: A robust benchmark and performant reward models. *arXiv preprint arXiv:2605.07872*, 2026.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

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

Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. From indoor to open world: Revealing the spatial reasoning gap in MLLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hou Xia, Zheren Fu, Fangcan Ling, Jiajun Li, Yi Tu, Zhendong Mao, and Yongdong Zhang. Video-LevelGauge: Investigating contextual positional bias in large video language models. In *International Conference on Learning Representations*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. SpatialTree: How spatial abilities branch out in MLLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, et al. ExpVid: A benchmark for experiment video understanding & reasoning. *arXiv preprint arXiv:2510.11606*, 2025.

Zhaoyang Xu, Xusheng He, Wei Liu, Zhenyang Li, and Jianlong Wu. Timelogic challenge@ cvpr 2026: Strong mllms meet evidence-seeking agents for temporal-logic video question answering. *arXiv preprint arXiv:2606.01631*, 2026a.

Zhiyang Xu, Tian Qin, Bowen Jin, Zhengfeng Lai, Meng Cao, Lifu Huang, and Peng Zhang. Incentivizing temporal-awareness in egocentric video understanding models. *arXiv preprint arXiv:2603.27184*, 2026b.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. In *Advances in Neural Information Processing Systems*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026a.

Yudong Yang, Guangzhi Sun, Yixuan Li, and Chao Zhang. SLVMBench: Skill learning from video memory. *arXiv preprint arXiv:2607.11312*, 2026b.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, et al. LongVT: Incentivizing "thinking with long videos" via native tool calling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026c.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591. IEEE, 2025.

Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin Zhang, Oluwatumininu Oguntola, and Gedas Bertasius. ExAct: A video-language benchmark for expert action analysis. In *Advances in Neural Information Processing Systems*, 2025.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, et al. Benchmarking visual state tracking in multimodal video understanding. *arXiv preprint arXiv:2606.03920*, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI Conference on Artificial Intelligence*, 2019.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing LMMs over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024a.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, et al. Task Me Anything. In *Advances in Neural Information Processing Systems*, 2024b.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b.

Le Zhang, Jihan Yang, Soundarya Krishnan, Jimit Majmudar, Xiou Ge, et al. From where things are to what they are for: Benchmarking spatial-functional intelligence in multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026b.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. ReVSI: Rebuilding visual spatial intelligence evaluation for accurate assessment of VLM 3D reasoning. In *International Conference on Machine Learning*, 2026c.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. *URL https://arxiv. org/abs/2410.02713*, 17(2):6, 2024c.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *International Conference on Computer Vision*, 2025c.

Zheyuan Zhang, Monica Dou, Linkai Peng, Hongyi Pan, Ulas Bagci, and Boqing Gong. VideoAds for fast-paced video understanding. In *International Conference on Computer Vision*, 2025d.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. DSI-Bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025e.

Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, et al. Q-Bench-Video: Benchmarking the video quality understanding of LMMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025f.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, et al. UrbanVideo-Bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In *Annual Meeting of the Association for Computational Linguistics*, 2025a.

Xinyi Zhao, Congjing Zhang, Pei Guo, Wei Li, Lin Chen, Chaoyue Zhao, and Shuai Huang. SmartHome-Bench: A comprehensive benchmark for video anomaly detection in smart homes using multi-modal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025b.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, et al. MMVU: Measuring expert-level multi-discipline video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c.

Yiming Zhao, Yu Zeng, Yukun Qi, YaoYang Liu, Xikun Bao, et al. V2P-Bench: Evaluating video-language understanding with visual prompts for better human-model interaction. In *International Conference on Learning Representations*, 2026.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, et al. Needle in a video haystack: A scalable synthetic evaluator for video MLLMs. In *International Conference on Learning Representations*, 2025d.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024a.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. *arXiv preprint arXiv:2210.15809*, 2022.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. ContPhy: Continuum physical concept learning and reasoning from videos. In *International Conference on Machine Learning*, 2024b.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, et al. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Qiji Zhou, Yifan Gong, Guangsheng Bao, Hongjie Qiu, Jinqiang Li, Xiangrong Zhu, Huajian Zhang, and Yue Zhang. Reasoning is all you need for video generalization: A counterfactual benchmark with sub-question evaluation. In *Findings of the Association for Computational Linguistics*, 2025b.

Sheng Zhou, Junbin Xiao, Xun Yang, Peipei Song, Dan Guo, Angela Yao, Meng Wang, and Tat-Seng Chua. Scene-text grounding for text-based video question answering. *IEEE Transactions on Multimedia*, 2025c.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4D: Dynamic spatial understanding for vision language models. *arXiv preprint arXiv:2512.20557*, 2025d.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, et al. MMR-V: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In *International Conference on Learning Representations*, 2026.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. VAU-R1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025.

Orr Zohar, Xiaohan Wang, Rui Li, Andrés Marafioti, Miquel Farré, et al. Apollo2: Exploring the long-video frontier of large multimodal models. `https://huggingface.co/datasets/Apollo-LMMs/LongTimeScope2`, 2025a.

Orr Zohar, Xiaohan Wang, Rui Li, Andrés Marafioti, Miquel Farré, et al. TimeScope (dataset). `https://huggingface.co/datasets/Apollo-LMMs/TimeScope`, 2025b.

Heqing Zou, Tianze Luo, Guiyang Xie, Victor Xiao Jie Zhang, Fengmao Lv, et al. HLV-1K: A large-scale hour-long video benchmark for time-specific long video understanding. In *IEEE International Conference on Multimedia and Expo*, 2025.

Zhishan Zou, Guoyan Sun, Zhiwei Wei, Jiancheng Pan, Yujie Li, Mugen Peng, and Wenjia Xu. Self in space: Benchmarking self-awareness and spatial cognition in UAV embodied intelligence. *arXiv preprint arXiv:2607.12477*, 2026.

APPENDIX

The appendix is organized as follows.

- Appendix A lists the 115 audited benchmarks with their references.
- Appendix B places the pyramid in the literature on benchmark validity, blind baselines, dataset artifacts, evidence grounding and evidence-seeking agents, data contamination, and benchmark composition and selection, and compares related audits in one table.
- Appendix C states the limitations of the audit and of Video-Index.
- Appendix D shows one item for each shortcut and two Video-Index items.
- Appendix E reads the five levels as bounded observers and states what each level can and cannot certify.
- Appendix F gives the per-benchmark figures of the option and text levels.
- Appendix G lists the settings of every stage of the audit, the pool, and Video-Index in one table.
- Appendix H details the measurement screen, the scoring rule, the error-attribution protocol, and the duplicate refill.
- Appendix I details the option level, with answer-position entropy, option counts against declared chance, and permutation sensitivity.
- Appendix J details the text level, with the blind readers, video dependence, and the paired McNemar test.
- Appendix K details the pool level, with shared videos and questions across benchmarks and the copy, retrieval, and learned attackers.
- Appendix L details the frame level, with the single-frame and caption attackers and the diagnostic profiles.
- Appendix M details the order level, with the shuffle, window, and half probes and the frame-budget ladder.
- Appendix N resolves the capability groups into fine categories and reports breaking levels by fine capability.
- Appendix O recounts the breaking levels with only the attacks that match the claim of each benchmark.
- Appendix P recounts the breaking levels at tolerances of 3% and 8% and with Claude Opus 5 as the reference, and compares the order in which benchmarks break under these settings.
- Appendix Q compares resolution against frame rate and reports the long-video frame-budget results.
- Appendix R shows where adding video lowers a score that the blind reader had right.
- Appendix S attributes the reference errors on surviving benchmarks to fine capability categories.
- Appendix T recommends benchmarks for each capability group.
- Appendix U details the pool construction, the watcher that keeps it current, and the Video-Index selection.
- Appendix V summarizes how the agents inspect video on Video-Index.
- Appendix W lists the taxonomy and the master table.
- Appendix X collects the report card of every audited benchmark.

A AUDITED BENCHMARK SOURCES

We evaluate the following 115 benchmarks. References are collected here so that the appendix tables can use benchmark names without repeated citations, namely ALLVB (Tan et al., 2025), AoTBench (Xue et al., 2025), Artifact-Bench (Tang et al., 2026a), Causal2Needles (Li et al., 2025),

CAViAR (Garg et al., 2026), CG-Bench (Chen et al., 2025b), CineTechBench (Wang et al., 2025c), CMU (Tan et al., 2026), Colon-Bench (Hamdi et al., 2026), ContPhy (Zheng et al., 2024b), COVER (Zhou et al., 2025b), DSI-Bench (Zhang et al., 2025e), DSR-Bench (Zhou et al., 2025d), E.T. Bench (Liu et al., 2024b), EgoPlan-Bench (Chen et al., 2025c), EgoSchema (Mangalam et al., 2023), Ex-Act (Yi et al., 2025), ExpVid (Xu et al., 2025), FAVOR-Bench (Tu et al., 2025), FineBench (Faure et al., 2026), ForeSeaQA (Park et al., 2026a), GasVideo-1000 (Tang et al., 2026b), GUI-World (Chen et al., 2025a), HERBench (Ben-Ami et al., 2026b), HLV-1K (Zou et al., 2025), HourVideo (Chandrasegaran et al., 2024), IntPhys 2 (Bordes et al., 2025), iSafetyBench (Abdullah et al., 2025), K9-Bench (Attarde et al., 2026), KFS-Bench (Li et al., 2026f), LEMONADE (Bonnetto et al., 2025), LongTimeScope (Zohar et al., 2025a), LongVideo-Reason (Chen et al., 2025d), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2024a), M3-Verse (Wei et al., 2025), MLVU (Zhou et al., 2025a), MME-VideoOCR (Shi et al., 2025), MMR-V (Zhu et al., 2026), MMSI-Video-Bench (Lin et al., 2025), MMVU (Zhao et al., 2025c), MMWorld (He et al., 2024), MORSE-500 (Cai et al., 2025), MotionBench (Hong et al., 2025), MVBench (Li et al., 2024b), NaturalBench-Video (Li, 2025), Neptune (Nagrani et al., 2024), NExT-QA (Xiao et al., 2021), OSI-Bench (Wu et al., 2026), OVO-S-Bench (Li et al., 2026e), PAI-Bench-U (Zhou et al., 2026), Perception Test (Patraucean et al., 2023), PerceptionComp (Li et al., 2026c), PhysicalAI (Zhou et al., 2026), PLM-VideoBench (Cho et al., 2025), Q-Bench-Video (Zhang et al., 2025f), ReactHuman (ReactHuman Benchmark Team, 2026), ReVSI (Zhang et al., 2026c), ReXTime (Chen et al., 2024a), S-EMBER (Wang et al., 2026c), SAW-Bench (Li et al., 2026b), ScaleLong (Ma et al., 2025), SciVideoBench (Deng et al., 2025), SFI-Bench (Zhang et al., 2026a), SIS-Bench (Zou et al., 2026), SITE-Bench (Wang et al., 2025b), SLVMBench (Yang et al., 2026b), SmartHome-Bench (Zhao et al., 2025b), SoMi-ToM (Fan et al., 2025), SpatialTree (Xiao et al., 2026), ST-Evidence (Wang et al., 2026b), STI-Bench (Li et al., 2025), StreamingBench (Lin et al., 2024), TaskMeAnything-v1 (Zhang et al., 2024b), TempCompass (Liu et al., 2024c), TemporalBench (Cai et al., 2024), TimeBlind (Li et al., 2026a), TimeScope (Zohar et al., 2025b), TOC-Bench (TOC-Bench Team, 2026), TOMATO (Shangguan et al., 2024), TVBench (Cores et al., 2024), UCF101-AD (Abdullah et al., 2026), UrbanVideo-Bench (Zhao et al., 2025a), v-HUB (Shi et al., 2026b), V2P-Bench (Zhao et al., 2026), VANE-Bench (Bharadwaj et al., 2025), VAU-Bench (Zhu et al., 2025), VCR-Bench (Qi et al., 2025), VE-LOCITI (Saravanan et al., 2025), VF-Eval (Song et al., 2025), VGenST-Bench (Park et al., 2026b), Video-Holmes (Cheng et al., 2025), Video-LevelGauge (Xia et al., 2026), Video-MME (Fu et al., 2025), Video-MME-v2 (Fu et al., 2026a), Video-MMMU (Hu et al., 2026), Video-TT (Zhang et al., 2025c), VideoAds (Zhang et al., 2025d), VideoEspresso (Han et al., 2025), VideoEval-Pro (Ma et al., 2026), VideoKR (Fu et al., 2026b), VideoMathQA (Rasheed et al., 2025), VideoOdyssey (He et al., 2026), VideoSIAH-Eval (Yang et al., 2026c), VidPair-Halluc (Huang et al., 2026a), ViMUL-Bench (Shafique et al., 2025), Vinoground (Zhang et al., 2024a), VITATECS (Li et al., 2024d), VKnowU (Jiang et al., 2025), VNBench (Zhao et al., 2025d), VRU-Accident (Kim et al., 2025), VSI-Bench (Yang et al., 2025b), VSI-Super-Wild (Gu et al., 2026), VSTAT (Yu et al., 2026), and VURB (Wei et al., 2026).

B Related Work

Video understanding benchmarks. ActivityNet-QA (Yu et al., 2019) and NExT-QA (Xiao et al., 2021) evaluate short clips. MVBench (Li et al., 2024b), Perception Test (Patraucean et al., 2023), EgoSchema (Mangalam et al., 2023), and Video-MME (Fu et al., 2025) broaden tasks and durations. Later evaluations focus on particular capabilities. HourVideo (Chandrasegaran et al., 2024), LVBench (Wang et al., 2024a), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025a) target long videos. TVBench (Cores et al., 2024), TOMATO (Shangguan et al., 2024), TempCompass (Liu et al., 2024c), and Vinoground (Zhang et al., 2024a) test temporal order. VSI-Bench (Yang et al., 2025b) evaluates spatial reasoning, VideoHallucer (Wang et al., 2024b) hallucination, and StreamingBench (Lin et al., 2024) online understanding. Audits reveal shortcuts within these settings. Buch et al. (2022) answer many earlier benchmark items from one selected frame. Lei et al. (2023) show that a single-frame model can outperform multi-frame methods on popular video-and-language datasets. TVBench and TOMATO show that shuffling can preserve temporal-benchmark scores. VideoEval-Pro (Ma et al., 2026) finds that guessing and priors inflate multiple-choice scores on long-video benchmarks and adopts open-ended short answers. Cambrian-S (Yang et al., 2026a) tests blind and single-frame performance and constructs VSI-Super (Yang et al., 2026a) to resist both. VBenchComp (Feng et al., 2025) separates text-answerable, semantic, and temporal items. Brown et al. (2025) train on question text with cross-validation and remove predictable items from

four test sets. We extend shortcut auditing to five levels across 115 benchmarks and compose new evaluations from surviving items.

Benchmark construction and auditing. MMStar (Chen et al., 2024b) removes image-benchmark items answerable without images or exposed through training data. NaturalBench (Li et al., 2024a) pairs each question with two images that yield different answers, which blocks blind solutions by design. DatBench (Joshi et al., 2026) selects items requiring images and separating models, while PerceptionBench (Lin et al., 2026) isolates atomic perception. BetterBench (Reuel et al., 2024) assesses evaluation practice. VLMEvalKit (Duan et al., 2024) and LMMs-Eval (Zhang et al., 2025b) run many benchmarks under shared protocols. tinyBenchmarks (Polo et al., 2024) and metabench (Kipnis et al., 2025) select items predictive of full scores. Fluid Benchmarking (Hofmann et al., 2025) selects adaptively. Lifelong Benchmarks (Prabhu et al., 2024), LiveXiv (Shabtay et al., 2025), and LiveBench (White et al., 2025) refresh evaluations over time. Perlitz et al. (2024) validate new benchmarks through agreement with established ones. Task-Me-Anything (Zhang et al., 2024b) generates items from taxonomies, AutoBencher (LI et al., 2025) searches for difficult topics, and BenchAgents (Butt et al., 2024) connects planning, generation, and verification. Harbor-Index (Shi et al., 2026a) distills 82 agentic tasks from 54 benchmarks. SWE-Bench+ (Aleithan et al., 2024) and BenchJack (Wang et al., 2026a) audit leaked solutions and exploitable graders. Our selection instruments build on AFLite's removal of predictable items (Le Bras et al., 2020), confident learning's label-error checks (Northcutt et al., 2021), and IFD's model-relative difficulty (Li et al., 2024c). Coverage-centric coreset selection (Zheng et al., 2022) and QDIT (Bukharin et al., 2024) balance coverage, quality, and diversity. We apply these principles to an audited pool drawn from existing video benchmarks. The selector preserves coverage, the Vendi score measures diversity (Friedman & Dieng, 2022), and a lockfile makes each composition reproducible.

Dataset artifacts and adversarial data. Language priors let visual question answering models answer without looking at the image, which motivated balanced image pairs in VQA v2 (Goyal et al., 2017) and shifted answer priors in VQA-CP (Agrawal et al., 2018). Hypothesis-only baselines expose annotation artifacts in natural language inference (Gururangan et al., 2018; Poliak et al., 2018). Adversarial collection with humans and models in the loop builds harder data (Nie et al., 2020; Kiela et al., 2021), yet filtering against one model family can make the resulting set unfair to that family (Phang et al., 2021). The attack pyramid extends the blind baseline to five levels of access to a video, and Qwen3-VL-8B (Bai et al., 2025), which screens the pool, is excluded from the ranking-agreement statistics of Video-Index. As shown in Table 2, prior audits apply one or two probes to a few benchmarks, while the pyramid applies five probes to 115 benchmarks and releases a screened pool for new compositions.

Evidence grounding and evidence seeking. Evidence-grounded benchmarks score whether a model finds the evidence behind its answer. EG-VQA (Huang et al., 2026b) pairs open-ended answers with annotated temporal evidence, and scene-text grounding (Zhou et al., 2025c) asks for the text regions that support each answer. ViTexQA (Guo et al., 2026) requires reading text across several frames. These benchmarks check the evidence directly, while Video-Index keeps the multiple-choice format and removes the items that attackers answer without it, and the two checks complement each other. Evidence-seeking systems choose frames before answering. Deep Video Discovery (Zhang et al., 2026b), TimeSearch-R (Pan et al., 2026), and T* (Ye et al., 2025) search the timeline for evidence. HiMu (Ben-Ami et al., 2026a) and ToolMerge (Shlapentokh-Rothman et al., 2026) decompose the question and score frames with lightweight tools, and an evidence-seeking agent for the TimeLogic challenge (Xu et al., 2026a) compares timestamped observations. Such searchers could serve as stronger attackers at the frame and order levels. UCS-Bench (Wang et al., 2026d) and EgoMonth (Chen et al., 2026) test memory over hours to months of egocentric video. TGPO (Xu et al., 2026b) rewards reasoning that depends on frame order by contrasting ordered and shuffled frames, the perturbation of our order level.

C LIMITATIONS

Multiple choice only. We audit only multiple-choice benchmarks. The protocol covers 115 English multiple-choice benchmarks, and open-ended and generative video benchmarks fall outside it. Video-Index keeps the multiple-choice format that the option level shows can leak answers, and its blind scores average four option orders.

Table 2: Related audits and benchmark builders. Data gives the modality, and Audit gives the number of existing benchmarks each work examines, where Several means standard benchmarks without a count in the abstract. Probes marks the shortcuts a work tests for or designs against, O for options alone, T for question text without the visual input, P for other evaluation items or training on the test set, F for single or few frames, and R for frame order.

Work	Data	Audit	Probes	Release
ATP (Buch et al., 2022)	Video	Several	F	Temporally hard subsets
Lei et al. (2023)	Video	Several	F	Two temporal retrieval tasks
TVBench (Cores et al., 2024)	Video	Several	T F	New benchmark
TOMATO (Shangguan et al., 2024)	Video	Several	F R	New benchmark, human annotated
VideoEval-Pro (Ma et al., 2026)	Video	Several	T F	Open-ended benchmark
TsT (Brown et al., 2025)	Image, video	4	T P	Debiased VSI-Bench
MMStar (Chen et al., 2024b)	Image	6	T	Curated subset, human reviewed
NaturalBench (Li et al., 2024a)	Image	–	T	Paired benchmark, human verified
tinyBenchmarks (Polo et al., 2024)	Text	4	–	Small predictive subsets
AutoBencher (LI et al., 2025)	Text	–	–	Generated datasets
This work	Video	115	O T P F R	Screened pool, compositions on request, Video-Index

No audio. We exclude benchmarks that require audio. No attacker receives speech, subtitles, or sound, and audio-visual claims remain untested.

Lower bounds. We bound exploitability from below. A benchmark that survives every level resists the tested attackers, and a stronger or newer attacker can raise its exploitability and lower its breaking level.

D QUALITATIVE EXAMPLES

We show one sampled item for each of five shortcuts and two items of Video-Index. Each box gives the question and the options, with the correct option in bold, and the text above it states what each attacker saw and chose. Frames appear where an attacker saw visual input or where the item belongs to Video-Index, and no attacker sees the correct answer.

Options alone. ExAct (Yi et al., 2025) breaks at the option level. The options-only attacker, Claude Opus 5 (Anthropic, 2026b), sees the five options of item 3175 without the question or the video and picks D, the correct option. D is the only option that says participant instead of player, and the only one without a reason clause. On 240 of the 300 sampled ExAct items, this attacker picks the correct option.

Question Which expert commentary correctly assesses the skill execution shown in the video?
A. The player should focus on jumping higher before releasing the ball to generate more backspin, improving the shot's precision and reducing bounce-off potential.
B. The player should focus on holding the elbow away from the body to create a wider arc, which will make the shot more difficult for opposing players to contest.
C. The player should keep the fingers stiff and straight while shooting to ensure a stronger shot and prevent the ball from being underpowered or easily blocked.
D. The participant should focus on snapping the wrist down and pointing the fingers down when shooting.
E. The player should concentrate on flicking the wrist upwards and extending the fingers straight out when shooting to achieve more height and distance on the shot.

Question text. MMR-V (Zhu et al., 2026) breaks at the text level. Gemini 3.5 Flash (Google DeepMind, 2026b) and Claude Opus 5 each read the question and the eleven options of item 337 without the video, and both pick D. Only D names a specific work, while the other ten options are generic titles about ratings, technology, and the workplace. Both readers answer 124 of the 300 sampled MMR-V items correctly without the video.

Question Please suggest a suitable title for this video.
A. A Digital Society Perspective B. Social Ratings and Personal Perceptions C. Online Influence on Personal Life **D. Black Mirror - Social Credit System** E. Living Under the Social Lens F. Empathy in the Age of Technology G. Corporate Culture vs. Personal Image H. The Intersection of Ratings and Reality I. Navigating Social Standards J. The Power of Community Approval K. Workplace Dynamics in a Digital World

Other evaluation items. TimeScope (Zohar et al., 2025b) breaks at the pool level. Items 1104 and 1892 come from videos of 3,600 and 180 seconds, yet they ask the same question with the same six options and the same answer. Their question embeddings coincide, and the copy attacker, which copies the answer of the nearest earlier question when the BGE-large (Xiao et al., 2024) cosine similarity exceeds 0.9, answers whichever item comes second.

Question What does the girl do after she get a cup of coffee?
A. reading a book B. working with her computer **C. taking some notes** D. cleaning the kitchen E. preparing a breakfast F. washing the dishes

Captions. NExT-QA (Xiao et al., 2021) breaks at the frame level. For item 7426, Claude Opus 5 reads the captions of 32 frames, each written independently by AuroraCap-7B-IMG (Chai et al., 2025), and picks C, the correct option. The captions describe a man with a small drone in a garden and later a small object resting on a white platform, and none of the 32 captions mentions a baby, a lady, or a girl. The same model shown a single frame picks B.



Question why did the blade stop moving at the end
A. baby stopped it B. no battery **C. the man stopped it** D. lady pressed a button E. girl pulled the sockets
Captions "A man in a green t-shirt and black shorts stands barefoot on a lawn, facing a small, black, circular object on the ground, which appears to be a remote-controlled toy." "The image shows a man standing in a garden, facing a drone that is in flight." "The image captures a tranquil garden scene with a small, unmanned aircraft on the ground, a white rectangular object on the grass, and a variety of greenery."

Shuffled frames. MVBench (Li et al., 2024b) breaks at the order level. For item action_prediction_88, Gemini 3.5 Flash reads the 32 frames in random order and still picks the correct next action, as it does with ordered frames. The question asks for the next action, yet the answer survives the loss of frame order.



Question What will the person do next?
A. Take the shoe. B. Lie on the floor. C. Close the box. **D. Tidy up the blanket.**

A Video-Index item that agents solve. NaturalBench-Video (Li, 2025) item 1481 plays an emergence in reverse, from a butterfly beside its chrysalis to the chrysalis alone. GPT-6-Astra (OpenAI, 2026), Claude Fable 5.1 (Anthropic, 2026a), and Claude Opus 5 working as agents pick B, the correct option. All nine open models and Claude Opus 5 with fixed input pick A, the forward transformation.



Question What transformation is occurring in the cocoon?
A. The insect in the cocoon turns into a butterfly. **B. The butterfly turns back into an insect in the cocoon.**

A Video-Index item that no system solves. S-EMBER (Wang et al., 2026c) item 3823 follows the wearer of a camera through a furniture store for about five minutes and asks how many price tags the wearer touches after closing a nightstand drawer. The answer is four. The three agents pick one, Claude Opus 5 with fixed input picks three, and all nine open models also miss the answer.

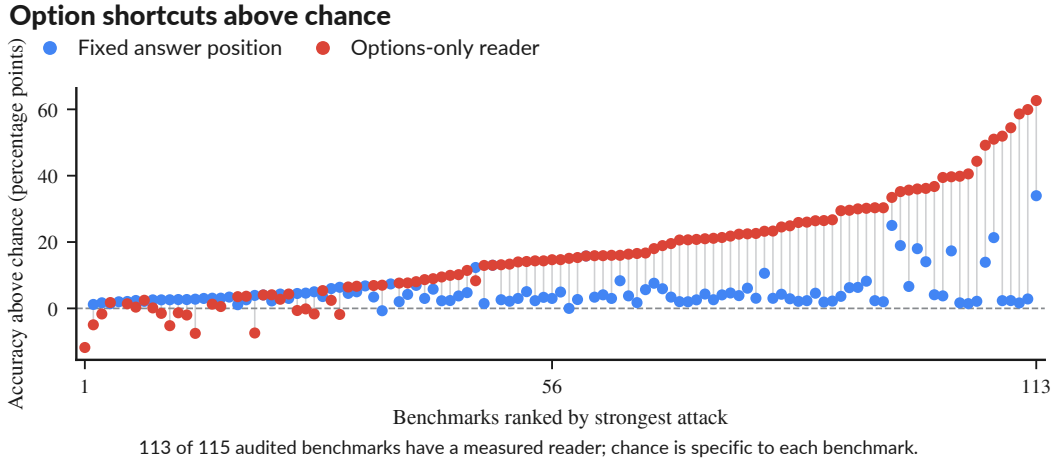


Question Since I closed the drawer on the nightstand, how many price tags have I touched?
A. 2 price tags. B. 3 price tags. C. 1 price tag. D. 0 price tags. **E. 4 price tags.**

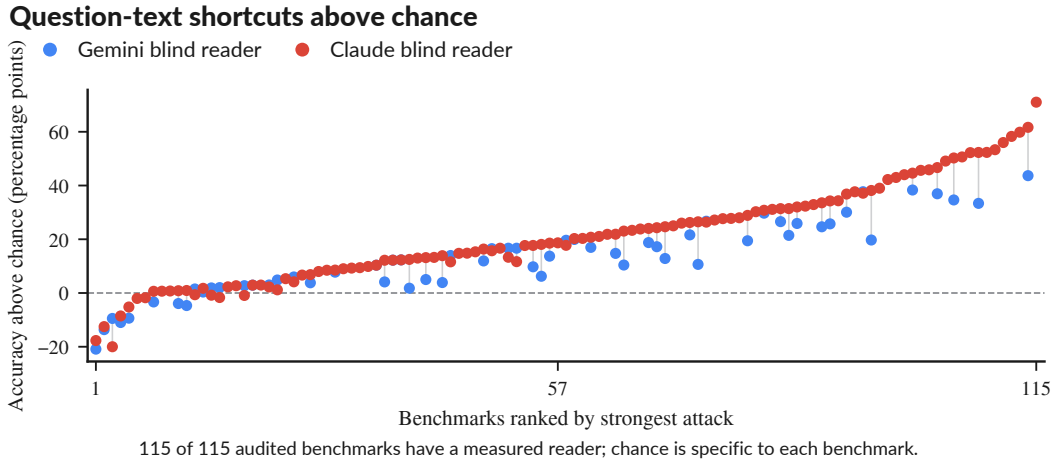
E Levels as Bounded Observers

Finzi et al. (2026) distinguish learnable structure from residual uncertainty relative to an observer's computational resources. Epiplexity denotes the structure available to a bounded observer. Time-bounded entropy denotes the remainder. This distinction motivates viewing each attacker set as an observer class. At the option level, the observer has only answer choices. Higher levels grant additional inputs. Exploitability ε_ℓ measures the score achieved by the declared class, and the breaking level identifies the first class approaching the reference. These operational measures differ from quantities such as answer-position entropy that depend only on the item distribution. A diverse-looking set can remain predictable to a shortcut model. At the pool level, sequential prediction measures how answer structure becomes learnable as earlier items accumulate.

Option permutations preserve the choices but change their positions. Frame shuffling preserves individual images but alters their order. Both interventions can change what a bounded model extracts



(a) Option attack margins above benchmark-specific chance across the audited multiple-choice sources.



(b) Blind-reader margins above benchmark-specific chance across the audited multiple-choice sources.

Figure F1: Measured option and blind-reader margins over chance within the 115-benchmark audit.

from the input. The former tests sensitivity to answer layout. The latter tests dependence on temporal order. We interpret their score changes relative to the stated capability and the information retained by each intervention. Evaluation aims to separate a capability-bearing model from shortcut observers. An informative item is difficult for the declared attackers but solvable from the intended evidence. The residual score above the strongest shortcut quantifies this separation. Admissibility is therefore relative to a claim and its tested threat model.

F PER-LEVEL FIGURES

As shown in Figure F1, option attacks and blind readers exceed each benchmark's own chance baseline on most audited benchmarks. Figure F1a compares fixed-position and options-only attacks, separating answer-position imbalance from predictability in the choices. Figure F1b compares two blind readers given questions and options. Each panel ranks benchmarks by its own stronger attack, and a horizontal position marks a different benchmark in each panel. Vertical lines connect paired measurements, showing how much the attackers differ on each benchmark. Zero denotes chance performance. A break requires approaching the full-video reference, and above-chance accuracy alone does not meet that criterion. Each panel holds the benchmarks with that measurement.

G REPRODUCIBILITY SETTINGS

Table 3 lists the settings of every stage of the audit, the screened pool, and Video-Index.

Table 3: Settings of the audit, the screened pool, and Video-Index. Each entry restates a setting given in the main text or the appendix.

Stage	Setting
Audit sample	300 questions per benchmark in a seed-42 draw stratified by subtask. One item per near-duplicate group is kept, and the draw refills within subtasks.
Media	Videos stored at a 720-pixel short side, or 480 for clips over 30 minutes without on-screen text. Frames reach the models at a long side of at most 768 pixels.
Reference	A reference model at 32 frames. Its 128-frame probe replaces it on long-video benchmarks where it stays within δ of chance.
Break rule	$\varepsilon_\ell \geq s^* - c - \delta$ with $\delta = 5\%$, about the 95% sampling margin of a 300-item accuracy.
Scoring	Option letter, then an answer-line regular expression, then the option text.
Option level	A fixed-position attacker picks the most frequent answer position, and Claude Opus 5 reads the options alone. Qwen3-VL-8B answers under eight option permutations with seed 42, 32 frames, and temperature 0.
Text level	Gemini 3.5 Flash and Claude Opus 5 read the question and options.
Pool level	A copy attacker at BGE-large cosine similarity above 0.9, Qwen3-8B reading the eight nearest question–answer pairs, and an online logistic regression on BGE-large embeddings over five rotating splits.
Frame level	Gemini 3.5 Flash reads one frame. Claude Opus 5 reads AuroraCap-7B-IMG captions of the 32 sampled frames, one caption per frame.
Order level	Gemini 3.5 Flash and Qwen3-VL-8B read 32 shuffled frames and, separately, a contiguous tenth of the video.
Duplicates	Near-duplicate questions at BGE-large cosine similarity of at least 0.9, confirmed with MPNet. Effective size is the Vendi score of 200 items under the product of SigLIP2 frame and BGE-large question similarities.
Pool screening	Qwen3-VL-8B answers from text alone, options alone, or one frame, and Qwen3-VL-2B from 32 frames in order or shuffled.
Composition	Round-robin selection over duration groups, event types, and sources with per-source caps. The gate reruns all five levels with Qwen3-VL-8B at 32 frames and the caption attacker and refills for a bounded number of rounds. A JSON lockfile records the snapshot hash, query, selector version, seed, item identifiers, option permutations, round count, and residual exploitability.
Video-Index	The 3,000 candidates with the lowest worst-case attacker percentile, one per video. Claude Opus 5 verifies answers at 32 frames, then at 512 frames, and 210 items fill each capability group.
Fixed-input evaluation	One frame per second, at most 512 frames, and a 224-pixel short side. Blind accuracy averages four option permutations with seed 42.
Agents	Local tools on videos stored at 2 fps and at most 1,024 frames.

H MEASUREMENT SCREEN

Scoring and error verification. The scorer applies the three-step rule in Section 3.1, checking a leading letter, a stated `Answer: C`, and option text. We verify errors through a reasoning-trace rerun on identical frames, rule-based pre-classification, two independent judges, and a dense-frame evidence check. Each judge, including Claude Opus 5 (Anthropic, 2026b), receives the same 32 hash-verified frames, question, options, gold answer, and incorrect answer with its reasoning. The responding model's identity is hidden. The checks distinguish annotation issues, insufficient frame coverage, and capability failures. Dense verification uses $\min(n_{\text{frames}}, 128)$ frames, where n_{frames} is the stored count. A self-check shifts the frame grid by half an interval on a tenth of confirmed failures. Judge disagreement marks a disputed case, an offset-induced answer flip an unstable case, and a missing dense check an unverified case. The remaining cases form the denominator for verified error shares.

Capability attribution. The profiles in Section 4.4 use the same 18 fine categories as the item taxonomy in Section 3.1. These comprise nine perception categories, four temporal categories,

three spatial and physical categories, and two reasoning and knowledge categories. Only cases with agreement on a fine label within this set contribute to Figure 10. Each share divides that category's count by the total across the 18 categories within the benchmark. The detailed tables retain other diagnostic labels, unresolved cases, and their original denominators.

Deduplication and replacement. The screen removes within-source near-duplicates and refills the audit sample toward $n = 300$. Duplicate questions share a video hash, options, and answer, with cosine similarity at least 0.90 under bge-large-en-v1.5 (Xiao et al., 2024). We retain one item per connected component. Refill continues the seed-42 stratified draw, replacing items within their subtasks before filling residual gaps in the original proportions. It excludes duplicates and non-multiple-choice items. A second screen checks scoring and duplication in the replacements.

Screening results. Table 4 reports error shares, including protocol failures and judge disagreement. Table 5 lists screening counts by benchmark. Across 115 benchmarks and 34,274 items, the screen removes 368 near-duplicates. Refill adds 287 items to the 39 benchmarks that fall below 300 after the screen. The frame-offset self-check flips 108 of 1,103 failures, or 9.8%, and the attribution excludes these as unstable. In Table 5, n sampled counts every sampled item of a listed benchmark, and pool remaining counts the never-sampled items of the full annotation pool after the audio-visual filter. Pool remaining counts items before the multiple-choice filter and the chance audit. Benchmarks without multiple-choice items are not listed. Every appendix table counts multiple-choice items only, and n is that count.

Table 4: Question-level error shares (%), including protocol failures and judge disagreement.

Benchmark	Percep.	Temporal	Spatial	Reason.	Proto.	Disput.	n
AoTBench	15.6	13.3	0.0	71.1	0.0	18.8	45
Artifact-Bench	46.6	0.0	6.9	46.6	2.0	20.1	149
CMU	35.2	0.0	5.6	59.3	1.8	25.0	55
COVER	26.9	3.9	3.9	65.4	30.7	27.3	75
DSI-Bench	7.7	0.0	65.4	26.9	0.9	2.5	105
DSR-Bench	9.2	20.2	29.4	41.2	4.0	22.7	124
EgoSchema	20.0	2.9	0.0	77.1	22.2	25.8	45
ExAct	26.2	0.0	0.0	73.8	6.2	17.2	65
ExpVid	56.4	9.1	0.0	34.5	75.2	11.0	222
FAVOR-Bench	40.8	12.7	0.0	46.5	10.1	19.0	79
GasVideo-1000	25.4	4.8	3.2	66.7	71.2	9.6	219
HourVideo	9.2	15.3	13.7	61.8	27.6	17.9	181
IntPhys 2	10.1	0.0	72.5	17.4	18.8	17.1	85
LEMONADE	0.0	78.9	16.9	4.2	68.9	11.3	228
LongTimeScope	50.0	0.0	0.0	50.0	87.6	2.1	194
LongVideoBench	10.5	15.8	5.3	68.4	58.7	8.1	46
LVBench	11.5	13.1	4.9	70.5	40.2	17.8	102
MLVU	26.9	3.9	0.0	69.2	59.4	10.7	64
MMR-V	2.8	1.4	0.0	95.8	25.0	26.3	96
MMSI-Video-Bench	7.0	3.1	49.2	40.6	14.1	9.2	149
MMVU	6.8	0.0	13.6	79.5	26.7	22.2	60
MMWorld	14.3	8.6	0.0	77.1	25.5	11.5	47
MORSE-500	19.4	29.6	22.4	28.6	15.5	24.8	116
MotionBench	30.0	16.7	5.0	48.3	9.1	9.8	66
MVBench	25.4	20.0	16.4	38.2	8.3	20.4	60
Neptune	33.3	33.3	0.0	33.3	72.7	15.4	11
OSI-Bench	1.5	7.5	70.7	20.3	3.6	7.4	138
Perception Test	31.9	8.5	4.3	55.3	4.1	28.1	49
PerceptionComp	27.1	0.7	39.4	32.9	16.7	17.7	186
Q-Bench-Video	47.6	0.0	0.0	52.4	38.2	23.3	34
ReVSI	16.9	0.0	62.1	21.1	11.7	0.0	188
SAW-Bench	2.1	0.0	58.7	39.2	0.0	6.5	143
SciVideoBench	0.0	57.1	0.0	42.9	84.3	21.5	223
SFI-Bench	25.0	0.0	16.7	58.3	3.2	19.4	62
SIS-Bench	79.3	0.9	4.3	15.5	55.4	7.3	260

Table 4 (continued)

Benchmark	Percep.	Temporal	Spatial	Reason.	Proto.	Disput.	n
SITE-Bench	23.7	6.8	30.5	39.0	9.2	10.2	65
SoMi-ToM	33.3	0.0	0.0	66.7	0.0	0.0	3
SpatialTree	14.6	0.0	52.1	33.3	15.8	6.5	57
ST-Evidence	27.3	9.1	4.5	59.1	15.4	23.5	26
STI-Bench	1.6	6.2	33.3	58.9	25.9	24.6	174
StreamingBench	16.3	30.2	2.3	51.2	30.6	21.1	62
TempCompass	17.0	5.1	3.4	74.6	56.6	8.0	136
TemporalBench	26.3	2.6	7.9	63.2	5.0	20.0	40
TimeBlind	28.1	12.5	6.2	53.1	5.9	19.1	34
TimeScope	30.0	0.0	0.0	70.0	82.5	5.7	114
TOMATO	22.2	22.2	13.3	42.2	10.0	29.5	100
TVBench	11.9	40.7	3.4	44.1	13.2	20.4	68
v-HUB	55.0	0.0	6.2	38.8	50.3	14.6	161
V2P-Bench	18.5	7.4	7.4	66.7	48.6	16.7	105
Video-Holmes	4.3	26.1	0.0	69.6	77.9	17.1	208
Video-MME	18.8	12.5	0.0	68.8	39.6	13.9	53
Video-MME-v2	7.6	15.2	3.8	73.4	28.2	15.7	110
Video-MMMU	10.0	3.3	3.3	83.3	37.5	25.6	48
Video-TT	16.7	25.0	0.0	58.3	14.3	22.2	14
VideoEval-Pro	13.4	6.0	1.5	79.1	46.8	6.3	126
VideoKR	12.8	0.0	5.1	82.0	51.8	30.5	81
VideoMathQA	16.7	0.0	12.5	70.8	42.9	11.8	42
VideoOdyssey	7.1	7.1	7.1	78.6	50.9	10.3	57
VideoSIAH-Eval	48.4	9.7	0.0	41.9	82.4	4.6	176
VidPair-Halluc	8.5	6.8	1.7	83.0	31.4	17.8	86
ViMUL-Bench	39.1	4.3	4.3	52.2	48.9	12.8	45
Vinoground	3.5	20.7	3.5	72.4	0.0	33.3	29
VITATECS	25.0	8.3	0.0	66.7	14.3	0.0	14
VKnowU	9.8	0.0	14.6	75.6	29.3	12.5	58
VSI-Bench	8.9	2.1	69.2	19.9	7.6	2.8	158
VSI-Super-Wild	20.2	24.6	0.0	55.3	14.9	16.5	134
VURB	19.4	13.9	16.7	50.0	41.9	22.6	62

Table 5: Audit items, near-duplicate removals, and replacement pools by benchmark.

Benchmark	n sampled	Dup. removed	n after	Pool remaining
CMU	300	0	300	3,853
COVER	300	1	299	5,546
DSI-Bench	299	0	299	1,470
DSR-Bench	299	0	299	1,151
EgoSchema	300	0	300	200
GasVideo-1000	650	72	578	363
LEMONADE	300	5	295	36,221
LongTimeScope	300	53	247	600
LongVideoBench	300	0	300	632
LVBench	300	1	299	1,249
MMWorld	300	1	299	1,259
MORSE-500	299	1	298	201
Neptune	300	19	281	8,500
OSI-Bench	299	0	299	8,467
PAI-Bench-U	300	1	299	914
Perception Test	300	1	299	18,840
SAW-Bench	299	10	289	3,843
SciVideoBench	300	0	300	700
SITE-Bench	300	0	300	3,319
SoMi-ToM	75	7	68	0

Table 5 (continued)

Benchmark	n sampled	Dup. removed	n after	Pool remaining
ST-Evidence	300	0	300	3,704
STI-Bench	299	13	286	3,829
TempCompass	300	9	291	7,240
TemporalBench	300	8	292	44,053
TimeBlind	300	1	299	2,100
TimeScope	300	14	286	2,700
TOMATO	300	0	300	1,184
TVBench	300	2	298	2,225
VCR-Bench	299	15	284	1,769
Video-Holmes	300	0	300	3,088
Video-MME-v2	300	0	300	2,590
Video-MMMU	300	0	300	900
Video-TT	300	6	294	9,700
VideoEval-Pro	300	1	299	989
VideoMathQA	300	0	300	1,800
VideoOdyssey	320	1	319	1,478
VidPair-Halluc	300	1	299	11,223
VKnowU	300	1	299	1,380
VSI-Bench	300	4	296	9,960
VSI-Super-Wild	300	3	297	6,680
VSTAT	344	6	338	1,156
VURB	300	76	224	1,800
Second run batch				
CG-Bench	120	0	120	12,009
CineTechBench	127	0	127	0
FineBench	300	22	278	19,843
ForeSeaQA	300	4	296	291
HERBench	300	1	299	1,671
iSafetyBench	299	0	299	3,157
K9-Bench	299	0	299	4,445
OVO-S-Bench	300	8	292	1,332
ReactHuman	297	0	297	757
S-EMBER	299	0	299	9,149
VAU-Bench	251	0	251	0
Displayed rows total	15,674	368	15,306	271,530

I OPTIONS WITHOUT QUESTIONS

Experiments. The option-level audit measures answer-position entropy, fixed-position accuracy, options-only accuracy, actual option counts, and sensitivity to option permutations. The final measure is the diagnostic model's best-to-worst accuracy range across reorderings.

Settings. We normalize ground-truth position entropy by the logarithm of the modal option count. The fixed-position attacker chooses the modal answer position. Claude Opus 5 (Anthropic, 2026b) receives `You will see ONLY the answer options of a multiple-choice question about a video.` A probe counts after covering 95% of items with options. Chance uses the reciprocal of the actual option count, then lettered choices in the question, the declared count, or a count inferred from answer letters. The audit reports disagreement with declared counts. Qwen3-VL-8B (Bai et al., 2025) evaluates eight permutations with seed 42, 32 frames, and temperature 0, and the original ordering is the ninth run. We flag accuracy ranges above $2.326\sqrt{c(1-c)/n}$, where c is chance accuracy and n the number of items. Item consistency is the fraction of runs choosing the same answer. Blind lift at or below -10% flags possible answer-format artifacts.

Results. Table 6 compares actual option counts and chance levels with their declared values.

Table 6: Declared option counts and chance levels that disagree with sampled benchmark items.

Benchmark	Decl. k	Modal k	Deviant %	Impl. chance	MCQ no ch.
ALLVB	–	5	0.0	0.200	yes
Artifact-Bench	NA	2	0.0	0.363	yes
Causal2Needles	–	4	0.0	0.250	yes
CAViAR	–	5	0.0	0.321	yes
CG-Bench	–	7	0.0	0.146	yes
CineTechBench	–	4	0.0	0.250	yes
CMU	4	2	56.3	0.367	no
ContPhy	–	2	0.0	0.396	yes
COVER	–	4	0.0	0.309	yes
E.T. Bench	–	4	0.0	0.233	yes
EgoPlan-Bench	4	4	0.0	0.250	yes
EgoSchema	4	5	100.0	0.200	no
ExpVid	NA	4	0.0	0.243	yes
FAVOR-Bench	–	5	0.0	0.200	yes
FineBench	–	4	0.0	0.250	yes
ForeSeaQA	–	4	0.0	0.334	yes
GasVideo-1000	NA	3	0.0	0.301	yes
GUI-World	–	4	0.0	0.250	yes
HERBench	–	5	0.0	0.200	yes
HLV-1K	–	4	0.0	0.250	yes
HourVideo	–	5	0.0	0.200	yes
IntPhys 2	–	2	0.0	0.500	yes
iSafetyBench	–	16	0.0	0.062	yes
K9-Bench	–	5	0.0	0.200	yes
KFS-Bench	–	5	0.0	0.211	yes
LongVideo-Reason	–	4	0.0	0.268	yes
LongVideoBench	5	5	27.3	0.214	yes
M3-Verse	–	4	0.0	0.236	yes
MLVU	–	4	0.0	0.250	yes
MME-VideoOCR	–	4	0.0	0.250	yes
MMR-V	NA	10	0.0	0.096	yes
MMSI-Video-Bench	–	4	0.0	0.242	yes
MMVU	–	5	0.0	0.200	yes
MVBench	–	4	0.0	0.287	yes
OVO-S-Bench	–	4	0.0	0.321	yes
Perception Test	–	3	0.0	0.333	yes
PLM-VideoBench	2	2	0.0	0.500	yes
ReactHuman	–	3	0.0	0.333	yes
ReXTime	–	4	0.0	0.250	yes
S-EMBER	–	5	0.0	0.200	yes
SAW-Bench	3	4	90.3	0.278	no
ScaleLong	–	4	0.0	0.250	yes
SciVideoBench	5	10	100.0	0.100	yes
SIS-Bench	–	4	0.0	0.250	yes
SITE-Bench	4	4	27.3	0.260	yes
SLVMBench	–	5	0.0	0.200	yes
SmartHome-Bench	–	2	0.0	0.500	yes
ST-Evidence	5	2	100.0	0.475	yes
STI-Bench	–	5	0.0	0.200	yes
StreamingBench	–	4	0.0	0.249	yes
TaskMeAnything-v1	–	4	0.0	0.253	yes
TOC-Bench	–	4	0.0	0.321	yes
TOMATO	4	6	83.0	0.199	no
TVBench	–	4	0.0	0.334	yes
UCF101-AD	–	11	0.0	0.091	yes
UrbanVideo-Bench	–	5	0.0	0.210	yes
VANE-Bench	–	4	0.0	0.250	yes

Table 6 (continued)

Benchmark	Decl. k	Modal k	Deviant %	Impl. chance	MCQ no ch.
VAU-Bench	–	4	0.0	0.250	yes
VELOCITI	–	2	0.0	0.500	yes
VF-Eval	–	4	0.0	0.250	yes
VGenST-Bench	–	4	0.0	0.267	yes
Video-Holmes	5	6	100.0	0.167	no
Video-LevelGauge	–	4	0.0	0.250	yes
Video-MME	4	4	0.0	0.250	yes
Video-MME-v2	5	8	100.0	0.127	yes
Video-MMMU	4	10	94.9	0.116	no
VideoAds	–	4	0.0	0.250	yes
VideoEspresso	–	4	0.0	0.250	yes
VideoEval-Pro	–	4	0.0	0.246	yes
VideoKR	NA	10	0.0	0.140	yes
VideoMathQA	5	2	81.7	0.445	yes
VidPair-Halluc	2	3	100.0	0.333	no
ViMUL-Bench	4	4	8.1	0.257	yes
Vinoground	–	2	0.0	0.500	yes
VITATECS	–	2	0.0	0.500	yes
VKnowU	NA	2	0.0	0.378	yes
VNBench	–	4	0.0	0.250	yes
VRU-Accident	–	4	0.0	0.250	yes
VSI-Super-Wild	NA	4	0.0	0.250	yes

J QUESTIONS WITHOUT VIDEO

Experiments. Gemini 3.5 Flash (Google DeepMind, 2026b) and Claude Opus 5 (Anthropic, 2026b) answer each question from text and options alone. Video dependence is the reference advantage over the stronger reader, $m = s^* - c - \varepsilon_{\text{text}}$. We apply a paired McNemar test to items answered both blind and with 32 frames by the same model. Fewer than 25 discordant pairs use an exact two-sided binomial test. Larger counts use a chi-square test. Tables 7–8 report condition accuracies, and Table 9 gives the paired tests.

Settings. Both readers receive independent calls without shared context. The prompt is `You will answer a question about a video WITHOUT seeing the video. Guess the most likely answer from the question and options alone.` It then supplies the question, lettered options, and a request for the option letter. Items without options request the exact short answer. Extraction skips Gemini's thinking text and applies the scorer in the appendix.

Results. Coverage is the share of multiple-choice items with a scored answer under the reference model's 32-frame condition.

Table 7: Input-condition accuracy (%) across audited multiple-choice benchmarks (1/2).

Benchmark	blind	1f	32f
ALLVB	73.3	70.3	–
AoTBench	49.1	49.8	76.5
Artifact-Bench	8.1	13.8	8.9
Causal2Needles	57.3	58.7	–
CAViAR	45.3	72.7	–
CG-Bench	47.5	51.7	–
CineTechBench	40.2	50.4	–
CMU	45.8	58.7	66.9
Colon-Bench	54.3	63.7	–
ContPhy	33.1	47.2	–

Table 7 (continued)

Benchmark	blind	1f	32f
COVER	57.2	56.8	59.2
DSI-Bench	51.2	53.5	52.7
DSR-Bench	24.1	30.8	23.9
E.T. Bench	28.7	32.7	–
EgoPlan-Bench	33.0	28.0	–
EgoSchema	56.7	62.7	77.0
ExAct	81.3	79.3	69.2
ExpVid	52.1	64.5	–
FAVOR-Bench	55.0	50.0	55.3
FineBench	63.3	57.4	–
ForeSeaQA	70.9	71.4	–
GasVideo-1000	44.9	57.0	75.4
GUI-World	81.0	82.7	–
HERBench	48.0	49.7	–
HLV-1K	67.0	65.5	–
HourVideo	22.7	20.3	21.2
IntPhys 2	48.7	47.7	57.0
iSafetyBench	14.7	36.1	–
K9-Bench	65.5	63.9	–
KFS-Bench	51.3	52.3	–
LEMONADE	41.7	32.3	–
LongTimeScope	29.7	36.3	15.7
LongVideo-Reason	76.3	73.2	–
LongVideoBench	52.7	56.3	60.4
LVBench	37.3	38.3	45.6
M3-Verse	26.3	33.7	–
MLVU	57.0	56.7	64.6
MME-VideoOCR	42.7	57.7	–
MMR-V	56.0	53.0	55.6
MMSI-Video-Bench	28.7	30.7	28.3
MMVU	57.5	78.7	76.9
MMWorld	77.8	83.5	74.4
MORSE-500	0.0	12.0	41.0
MotionBench	50.3	59.0	64.9
MVBench	46.0	53.7	68.9
NaturalBench-Video	50.7	54.7	–
Neptune	72.2	70.6	–
NExT-QA	59.0	75.3	–
OSI-Bench	38.5	36.3	47.8
OVO-S-Bench	40.4	41.8	–
PAI-Bench-U	34.2	49.7	–
Perception Test	45.0	62.0	74.7
PerceptionComp	7.4	8.4	13.5
PLM-VideoBench	71.7	82.0	–
Q-Bench-Video	48.1	63.2	72.6
ReactHuman	31.3	40.4	–
ReVSI	39.3	29.5	52.1
ReXTime	84.3	87.7	–
S-EMBER	29.4	21.7	–
SAW-Bench	28.0	25.4	38.1
ScaleLong	48.7	52.5	–
SciVideoBench	54.0	52.0	–
SFI-Bench	62.3	56.7	68.9
SIS-Bench	50.0	58.3	–
SITE-Bench	50.2	45.8	65.3
SLVMBench	63.0	64.7	–
SmartHome-Bench	63.0	71.3	–
SoMi-ToM	46.7	50.0	93.3
SpatialTree	69.2	62.5	62.2

Table 7 (continued)

Benchmark	blind	1f	32f
ST-Evidence	66.2	65.6	82.1
STI-Bench	47.2	49.2	–
StreamingBench	56.0	54.3	61.9
TaskMeAnything-v1	32.0	41.3	–
TempCompass	27.8	30.9	42.4
TemporalBench	76.4	75.0	77.0
TimeBlind	47.5	49.2	70.9
TimeScope	40.3	42.3	52.6
TOC-Bench	35.0	41.1	–
TOMATO	19.0	22.3	43.4
TVBench	40.7	44.0	62.8
UCF101-AD	7.3	35.7	–
UrbanVideo-Bench	41.3	41.3	–
v-HUB	20.8	56.1	64.3
V2P-Bench	41.0	46.7	52.7
VANE-Bench	77.3	87.3	–
VAU-Bench	96.0	94.8	–
VCR-Bench	45.1	47.0	–
VELOCITI	65.3	75.3	–
VF-Eval	83.3	81.3	–
VGenST-Bench	45.3	60.7	–
Video-Holmes	54.3	58.0	–
Video-LevelGauge	52.7	64.7	–
Video-MME	56.7	56.7	70.2
Video-MME-v2	29.3	28.0	36.4
Video-MMMU	56.2	59.2	70.5
Video-TT	40.0	40.0	60.7
VideoAds	51.0	57.3	–
VideoEspresso	75.0	79.3	–
VideoEval-Pro	35.3	37.3	45.0
VideoKR	35.1	43.2	55.2
VideoMathQA	57.3	63.0	51.7
VideoOdyssey	44.7	45.3	34.8
VideoSIAH-Eval	6.9	8.7	12.8
VidPair-Halluc	29.8	41.3	42.7
ViMUL-Bench	43.4	62.4	80.1
Vinoground	50.0	54.3	82.2
VITATECS	78.3	89.3	92.0
VKnowU	63.3	60.6	71.5
VNBench	26.0	25.4	–
VRU-Accident	48.3	67.3	–
VSI-Bench	39.0	29.4	60.8
VSI-Super-Wild	25.4	31.4	32.5
VSTAT	43.8	42.2	–
VURB	38.4	42.2	62.4

Table 8: Input-condition accuracy (%) across audited multiple-choice benchmarks (2/2).

Benchmark	32f	video cap.	Frame cap.	Cov.
ALLVB	72.7	71.0	72.0	1.00
AoTBench	80.0	72.2	70.5	0.98
Artifact-Bench	17.1	11.4	13.8	1.00
Causal2Needles	63.7	59.7	60.3	1.00
CAViAR	77.7	75.0	71.3	1.00
CG-Bench	56.7	47.5	52.5	1.00
CineTechBench	60.2	54.9	56.6	0.89

Table 8 (continued)

Benchmark	32f	video cap.	Frame cap.	Cov.
CMU	72.3	56.8	58.1	0.99
Colon-Bench	74.3	62.0	54.3	1.00
ContPhy	55.8	33.7	33.7	1.00
COVER	62.9	59.8	60.3	1.00
DSI-Bench	63.9	59.9	55.2	1.00
DSR-Bench	29.1	34.1	30.8	0.78
E.T. Bench	54.7	51.7	48.3	1.00
EgoPlan-Bench	29.3	30.0	32.7	1.00
EgoSchema	78.3	80.3	71.0	0.94
ExAct	83.7	79.0	82.0	0.96
ExpVid	64.5	62.7	59.0	1.00
FAVOR-Bench	67.0	54.3	52.3	0.98
FineBench	61.9	62.6	60.7	0.52
ForeSeaQA	73.4	68.8	69.3	1.00
GasVideo-1000	75.4	72.7	62.2	0.97
GUI-World	92.0	88.3	86.3	1.00
HERBench	55.7	54.7	54.3	1.00
HLV-1K	68.9	69.9	69.6	0.99
HourVideo	24.0	24.0	22.0	0.96
IntPhys 2	64.3	55.3	52.3	0.99
iSafetyBench	42.1	42.8	36.4	1.00
K9-Bench	74.8	68.7	67.0	0.98
KFS-Bench	67.3	61.7	60.0	1.00
LEMONADE	40.0	37.7	38.0	1.00
LongTimeScope	33.3	27.3	28.0	0.89
LongVideo-Reason	77.3	79.0	76.6	0.98
LongVideoBench	76.3	69.7	63.0	0.93
LVBench	58.7	47.0	46.7	0.95
M3-Verse	40.3	34.0	29.3	1.00
MLVU	74.7	68.7	66.3	0.93
MME-VideoOCR	84.0	67.7	52.3	1.00
MMR-V	61.0	58.4	57.0	0.92
MMSI-Video-Bench	35.3	35.0	27.0	0.92
MMVU	89.9	83.5	78.2	0.97
MMWorld	83.5	80.8	80.8	0.96
MORSE-500	38.0	24.1	10.2	0.88
MotionBench	69.3	63.7	56.3	0.99
MVBench	67.7	66.3	63.0	0.96
NaturalBench-Video	80.7	66.7	60.0	1.00
Neptune	81.3	75.6	–	–
NExT-QA	83.3	79.3	78.3	–
OSI-Bench	46.2	41.8	37.4	0.99
OVO-S-Bench	46.9	48.0	43.6	1.00
PAI-Bench-U	60.7	56.1	42.6	1.00
Perception Test	77.3	71.3	65.7	0.99
PerceptionComp	16.4	9.4	8.0	0.89
PLM-VideoBench	84.8	79.8	75.8	0.92
Q-Bench-Video	59.4	46.2	53.8	1.00
ReactHuman	50.2	51.2	45.8	1.00
ReVSI	54.1	40.2	34.4	0.98
ReXTime	91.8	89.8	91.0	0.81
S-EMBER	37.1	24.1	28.4	1.00
SAW-Bench	32.7	32.4	32.7	0.98
ScaleLong	64.0	59.9	57.6	0.99
SciVideoBench	65.0	62.0	57.7	1.00
SFI-Bench	74.3	60.7	61.7	0.99
SIS-Bench	72.0	69.0	61.7	1.00
SITE-Bench	68.1	64.9	59.9	0.99
SLVMBench	79.7	73.0	73.3	1.00

Table 8 (continued)

Benchmark	32f	video cap.	Frame cap.	Cov.
SmartHome-Bench	78.0	71.3	66.7	1.00
SoMi-ToM	100.0	85.0	70.0	0.80
SpatialTree	70.3	59.4	64.1	0.92
ST-Evidence	84.0	80.8	77.2	0.95
STI-Bench	48.8	48.2	44.5	1.00
StreamingBench	69.0	62.0	59.7	0.97
TaskMeAnything-v1	61.7	49.7	44.0	1.00
TempCompass	39.7	39.7	34.1	0.99
TemporalBench	79.6	76.1	78.5	0.99
TimeBlind	82.3	66.3	60.8	0.61
TimeScope	64.0	57.7	50.7	0.95
TOC-Bench	49.6	46.8	46.8	1.00
TOMATO	47.0	29.8	29.3	0.98
TVBench	66.0	56.3	49.7	0.99
UCF101-AD	46.7	48.7	34.3	1.00
UrbanVideo-Bench	54.0	44.0	49.0	1.00
v-HUB	74.6	70.8	67.7	0.97
V2P-Bench	58.0	53.0	54.7	0.98
VANE-Bench	88.3	82.0	81.0	1.00
VAU-Bench	94.4	95.2	96.4	1.00
VCR-Bench	66.3	57.7	51.7	0.99
VELOCITI	93.7	93.0	88.0	1.00
VF-Eval	84.0	81.7	79.3	1.00
VGenST-Bench	87.7	80.7	70.7	1.00
Video-Holmes	67.7	62.7	56.0	1.00
Video-LevelGauge	89.0	74.0	67.7	1.00
Video-MME	80.0	70.7	65.0	0.96
Video-MME-v2	48.0	39.0	35.3	0.84
Video-MMMU	80.6	69.6	54.9	0.91
Video-TT	60.0	45.0	35.0	0.93
VideoAds	81.3	75.3	76.3	1.00
VideoEspresso	79.6	76.7	77.8	0.92
VideoEval-Pro	56.7	46.3	47.7	0.96
VideoKR	67.0	56.8	44.9	0.88
VideoMathQA	80.7	69.0	59.0	0.69
VideoOdyssey	50.0	49.4	47.8	0.93
VideoSIAH-Eval	12.1	10.8	12.1	0.94
VidPair-Halluc	46.2	42.3	41.3	0.94
ViMUL-Bench	82.1	74.6	66.5	0.93
Vinoground	86.3	78.7	76.0	0.98
VITATECS	93.3	88.7	88.7	0.96
VKnowU	76.1	77.3	72.7	0.97
VNBench	73.9	60.2	51.8	1.00
VRU-Accident	76.0	74.7	64.7	1.00
VSI-Bench	59.6	44.5	53.4	0.89
VSI-Super-Wild	32.9	31.4	25.0	0.76
VSTAT	48.6	49.6	48.9	–
VURB	63.1	58.2	50.5	0.91

Table 9: Paired blind versus 32-frame McNemar tests, where b and c count discordant pairs.

Benchmark	n pairs	b	c	p
ALLVB	300	30	28	0.8955
AoTBench	290	21	106	0.0000
Artifact-Bench	123	6	10	0.4545
Causal2Needles	300	35	54	0.0564

Table 9 (continued)

Benchmark	n pairs	b	c	p
CAViAR	300	15	112	0.0000
CG-Bench	120	13	24	0.1002
CineTechBench	113	15	38	0.0025
CMU	154	18	61	0.0000
Colon-Bench	300	28	88	0.0000
ContPhy	163	19	56	0.0000
COVER	228	35	54	0.0564
DSI-Bench	298	56	59	0.8521
DSR-Bench	234	40	34	0.5611
E.T. Bench	300	31	109	0.0000
EgoPlan-Bench	100	25	20	0.5510
EgoSchema	282	15	120	0.0000
ExAct	289	26	42	0.0689
FAVOR-Bench	293	45	73	0.0129
FineBench	155	23	32	0.2807
ForeSeaQA	199	17	22	0.5218
GasVideo-1000	382	26	136	0.0000
GUI-World	300	6	39	0.0000
HERBench	300	31	54	0.0170
HLV-1K	296	33	38	0.6350
HourVideo	288	41	36	0.6485
IntPhys 2	298	103	118	0.3463
iSafetyBench	299	17	99	0.0000
K9-Bench	294	27	54	0.0039
KFS-Bench	300	29	77	0.0000
LongTimeScope	267	71	24	0.0000
LongVideo-Reason	295	24	28	0.6774
LongVideoBench	280	50	76	0.0259
LVBench	285	41	88	0.0001
M3-Verse	300	33	75	0.0001
MLVU	280	45	87	0.0004
MME-VideoOCR	300	11	135	0.0000
MMR-V	277	32	57	0.0110
MMSI-Video-Bench	276	57	52	0.7016
MMVU	182	19	54	0.0001
MMWorld	285	27	70	0.0000
MORSE-500	95	4	33	0.0000
MotionBench	296	30	109	0.0000
MVBench	289	26	114	0.0000
NaturalBench-Video	150	10	55	0.0000
OSI-Bench	90	7	24	0.0041
OVO-S-Bench	275	36	54	0.0731
Perception Test	296	25	106	0.0000
PerceptionComp	267	12	31	0.0061
PLM-VideoBench	277	16	52	0.0000
Q-Bench-Video	106	12	42	0.0001
ReactHuman	297	21	77	0.0000
ReVSI	118	12	36	0.0009
ReXTime	244	6	23	0.0030
S-EMBER	299	38	61	0.0270
SAW-Bench	270	45	70	0.0252
ScaleLong	297	27	72	0.0000
SFI-Bench	296	27	68	0.0000
SITE-Bench	297	44	110	0.0000
SLVMBench	300	18	68	0.0000
SmartHome-Bench	300	47	92	0.0002
SoMi-ToM	60	1	27	0.0000
SpatialTree	116	13	26	0.0547
ST-Evidence	268	21	77	0.0000

Table 9 (continued)

Benchmark	n pairs	b	c	p
StreamingBench	291	43	88	0.0001
TaskMeAnything-v1	300	19	108	0.0000
TempCompass	125	2	27	0.0000
TemporalBench	282	28	43	0.0966
TimeBlind	110	16	36	0.0084
TimeScope	285	53	102	0.0001
TOC-Bench	280	22	63	0.0000
TOMATO	295	31	95	0.0000
TVBench	296	31	107	0.0000
UCF101-AD	300	11	129	0.0000
UrbanVideo-Bench	300	26	64	0.0001
v-HUB	126	3	63	0.0000
V2P-Bench	294	38	84	0.0000
VANE-Bench	300	8	41	0.0000
VAU-Bench	251	9	5	0.4239
VELOCITI	300	13	98	0.0000
VF-Eval	299	23	26	0.7751
VGenST-Bench	300	9	136	0.0000
Video-LevelGauge	300	10	119	0.0000
Video-MME	289	31	85	0.0000
Video-MME-v2	253	25	69	0.0000
Video-MMMU	269	22	78	0.0000
Video-TT	56	3	22	0.0003
VideoAds	300	18	109	0.0000
VideoEspresso	275	19	30	0.1531
VideoEval-Pro	289	47	76	0.0116
VideoKR	161	15	64	0.0000
VideoMathQA	207	49	38	0.2837
VideoOdyssey	296	73	43	0.0071
VideoSIAH-Eval	219	7	26	0.0017
VidPair-Halluc	96	6	24	0.0019
ViMUL-Bench	161	13	69	0.0000
Vinoground	293	15	121	0.0000
VITATECS	288	11	76	0.0000
VKnowU	255	21	84	0.0000
VNBench	299	18	162	0.0000
VRU-Accident	300	19	102	0.0000
VSI-Bench	130	15	46	0.0001
VSI-Super-Wild	191	27	32	0.6025
VURB	189	16	61	0.0000

Table 10: Text-exposure indicators, namely blind lift, min-20%, zlib, and infini-gram matches.

Benchmark	Yr	Lift	mink 7B	mink 1.5B	mink Mistral	zlib	ngram %
VNBench	2025	1.0	6.99	6.37	6.04	0.148	16.0
v-HUB	2025	-3.1	9.82	8.45	7.78	0.186	21.7
SoMi-ToM	2025	17.3	9.69	8.85	8.56	0.195	0.0
VGenST-Bench	2026	18.7	8.01	7.56	6.94	0.156	10.7
MME-VideoOCR	2025	17.7	7.34	6.74	6.40	0.166	11.0
UCF101-AD	2026	-1.8	6.37	5.82	5.35	0.167	0.0
EgoSchema	2023	19.3	6.73	5.99	5.58	0.179	0.7
Video-LevelGauge	2026	27.7	7.96	7.48	6.70	0.174	18.0
Vinoground	2024	-4.0	7.53	6.69	6.35	0.177	0.0
ViMUL-Bench	2025	18.3	7.32	6.39	5.90	0.189	3.0
Video-TT	2025	3.3	7.68	6.76	6.36	0.173	2.3
CAViAR	2026	13.2	6.20	5.93	5.53	0.114	25.0

Table 10 (continued)

Benchmark	Yr	Lift	mink 7B	mink 1.5B	mink Mistral	zlib	ngram %
VideoKR	2026	9.1	8.03	7.48	7.03	0.169	7.7
MORSE-500	2025	-9.8	7.44	7.29	6.88	0.162	100.0
VideoAds	2025	26.0	6.31	5.78	5.47	0.140	19.7
NaturalBench-Video	2025	0.7	8.35	7.54	7.46	0.149	13.7
MVBench	2023	10.3	9.15	7.82	7.52	0.177	27.3
TaskMeAnything-v1	2024	6.7	8.47	7.28	6.93	0.157	10.7
GasVideo-1000	2026	16.1	11.02	8.49	8.50	0.189	16.8
AoTBench	2025	0.7	8.77	7.79	6.66	0.191	4.7
VELOCITI	2025	15.3	7.19	6.42	5.97	0.181	0.0
Q-Bench-Video	2024	17.0	7.79	7.29	7.11	0.145	10.0
VRU-Accident	2025	23.3	7.85	6.58	6.34	0.145	14.7
Perception Test	2023	13.7	9.04	7.61	7.37	0.188	25.3
iSafetyBench	2025	8.5	6.80	6.22	5.71	0.168	3.0
MotionBench	2025	13.0	6.88	6.39	6.24	0.151	13.7
PAI-Bench-U	2026	9.0	9.12	8.38	8.11	0.170	1.3
E.T. Bench	2024	5.3	8.08	6.93	6.36	0.187	31.7
TVBench	2024	4.0	8.25	7.19	6.76	0.173	34.0
VKnowU	2025	9.9	6.93	6.25	6.12	0.156	17.3
VSI-Bench	2024	7.5	7.94	6.66	6.58	0.150	1.3
VITATECS	2023	18.5	7.12	6.26	6.27	0.155	5.3
ContPhy	2024	2.3	8.63	7.84	7.58	0.169	8.3
VURB	2026	13.2	8.02	7.15	6.72	0.162	10.0
SITE-Bench	2025	17.3	9.07	7.62	7.32	0.185	21.0
Video-MMMU	2025	36.6	6.55	6.01	5.55	0.195	14.3
SIS-Bench	2026	25.0	9.16	8.43	7.77	0.146	0.0
TimeBlind	2026	-0.8	7.91	7.21	7.23	0.146	12.3
TOMATO	2024	2.1	9.36	8.18	7.00	0.189	8.0
VCR-Bench	2025	21.1	8.91	8.12	7.89	0.153	0.0
ReVSI	2026	4.3	8.23	7.18	7.15	0.158	4.0
ST-Evidence	2026	14.2	7.90	6.95	6.98	0.160	18.7
TempCompass	2024	-4.5	8.57	7.54	7.03	0.162	5.3
CineTechBench	2025	15.2	8.40	7.82	6.72	0.185	29.9
MMVU	2025	36.9	7.28	6.99	6.66	0.137	12.7
Colon-Bench	2026	34.3	6.76	6.68	5.86	0.163	12.7
OSI-Bench	2026	3.6	8.17	7.60	7.04	0.181	0.0
ReactHuman	2026	-2.0	8.41	7.06	6.71	0.151	100.0
Video-MME	2024	27.0	7.01	6.57	6.17	0.153	14.0
Video-MME-v2	2026	6.0	6.50	6.12	5.58	0.178	5.0
LVBench	2024	3.7	8.33	7.26	6.42	0.174	17.0
SLVMBench	2026	43.0	6.30	5.92	5.41	0.148	19.7
TimeScope	2025	19.3	9.71	7.57	6.71	0.199	57.7
MLVU	2024	23.5	7.79	7.04	6.29	0.152	16.7
KFS-Bench	2026	30.2	7.23	6.55	6.11	0.181	17.0
MMWorld	2024	32.7	7.58	7.00	6.40	0.167	9.0
VidPair-Halluc	2026	-14.2	9.00	8.13	7.77	0.181	0.0
V2P-Bench	2025	11.9	9.72	7.82	6.93	0.196	17.3
StreamingBench	2024	21.4	7.09	6.70	6.13	0.157	15.0
ScaleLong	2025	23.7	6.54	6.16	5.98	0.144	17.0
SmartHome-Bench	2025	13.0	5.87	6.13	5.57	0.168	0.0
TOC-Bench	2026	2.9	8.72	7.95	7.54	0.182	15.0
SFI-Bench	2026	29.7	7.57	6.85	6.29	0.164	20.0
M3-Verse	2025	2.7	7.81	7.19	7.28	0.183	9.7
Video-Holmes	2025	37.7	8.49	8.02	7.27	0.171	0.3
PLM-VideoBench	2025	21.7	6.33	5.78	5.47	0.152	33.3
UrbanVideo-Bench	2025	20.3	7.42	6.58	6.24	0.199	7.0
ExpVid	2025	27.8	8.70	7.97	7.41	0.169	6.3
SpatialTree	2025	33.3	7.42	6.53	5.95	0.146	1.3
GUI-World	2025	56.0	7.03	6.55	6.09	0.145	8.7
VANE-Bench	2025	52.3	6.87	6.74	6.34	0.141	17.0

Table 10 (continued)

Benchmark	Yr	Lift	mink 7B	mink 1.5B	mink Mistral	zlib	ngram %
SciVideoBench	2025	44.0	7.83	7.12	6.78	0.177	0.0
LongVideoBench	2024	29.3	8.24	7.21	6.47	0.199	14.7
VideoEval-Pro	2025	10.7	7.44	6.89	6.54	0.154	18.3
SAW-Bench	2026	2.9	8.77	7.84	7.22	0.189	0.3
K9-Bench	2026	45.5	7.51	6.90	6.12	0.201	0.0
CG-Bench	2024	32.9	6.36	6.06	5.84	0.163	16.7
VideoSIAH-Eval	2025	-21.1	10.54	9.37	8.42	0.195	15.0
FAVOR-Bench	2025	26.7	8.00	7.28	6.87	0.181	7.3
MMR-V	2025	37.4	7.50	6.81	5.86	0.165	7.0
COVER	2025	26.1	7.42	6.75	6.61	0.149	12.3
S-EMBER	2026	9.4	6.46	5.79	5.38	0.171	10.4
HERBench	2026	28.0	6.98	6.54	5.89	0.173	9.3
ReXTime	2024	59.3	6.96	6.40	5.96	0.160	5.0
PerceptionComp	2026	-13.3	8.71	8.02	7.21	0.194	8.3
OVO-S-Bench	2026	9.9	7.32	7.40	7.10	0.196	13.7
Causal2Needles	2025	32.3	9.04	8.12	7.21	0.189	12.3
IntPhys 2	2025	1.7	10.61	8.41	8.65	0.160	0.0
ExAct	2025	44.0	6.45	6.36	5.80	0.175	0.0
TemporalBench	2024	21.8	6.71	6.15	5.48	0.215	0.0
VideoEspresso	2025	50.0	6.82	6.47	5.80	0.172	0.0
VSI-Super-Wild	2026	4.0	7.56	7.30	6.86	0.174	0.0
Artifact-Bench	2026	-11.0	10.37	8.31	7.93	0.193	0.0
ForeSeaQA	2026	45.9	6.53	6.26	6.31	0.147	10.7
HLV-1K	2025	42.0	6.67	6.29	5.89	0.166	8.7
STI-Bench	2025	27.2	8.09	7.39	6.79	0.185	0.0
DSI-Bench	2025	26.5	6.71	6.41	6.25	0.154	0.0
LongVideo-Reason	2025	49.5	8.06	7.51	6.71	0.175	6.3
VF-Eval	2025	58.3	7.61	6.73	6.26	0.163	3.3
DSR-Bench	2025	-0.9	6.43	5.74	5.50	0.222	0.7
ALLVB	2025	53.3	6.73	6.48	5.93	0.148	16.0
FineBench	2026	38.3	7.47	6.98	6.59	0.193	20.7
HourVideo	2024	2.7	6.97	6.31	5.90	0.177	6.0
VAU-Bench	2025	71.0	6.30	6.11	5.85	0.138	0.0
LEMONADE	2025	16.7	10.67	9.26	8.11	0.190	3.7
MMSI-Video-Bench	2025	5.9	6.95	6.42	6.26	0.166	8.0
EgoPlan-Bench	2023	8.0	9.04	7.49	7.60	0.189	33.3
VideoMathQA	2025	11.7	7.31	6.18	6.44	0.166	9.3
VideoOdyssey	2026	19.4	7.12	6.63	6.22	0.164	10.9
LongTimeScope	2025	17.0	9.67	7.50	6.65	0.198	56.7
NExT-QA	2021	–	9.05	7.76	8.06	0.209	44.3
Neptune	2024	–	8.79	7.79	6.96	0.159	0.7
VSTAT	2026	–	8.66	7.72	7.10	0.171	4.1

Table 10 reports three text-exposure indicators, namely min-$k\%$ scores under three open-weight models (Shi et al., 2024), zlib compression ratios, and exact question matches in Dolma (Liu et al., 2024a; Soldaini et al., 2024). Across 110 benchmarks with both indicators, blind lift correlates with Qwen2.5-7B (Qwen Team, 2024) min-$k\%$ at Spearman -0.38. These indicators are indirect evidence of exposure. Median blind lift by release year is $+14\%$ for 2023, $+8\%$ for 2024, $+21\%$ for 2025, and $+13\%$ for 2026. A benchmark's release date need not date its source material, and a question match establishes public text rather than answer access. The text attack directly measures answerability without video.

K Learning from evaluation items

Shared content and duplicates. Content hashes identify 923 shared videos and 2,893 cross-benchmark video pairs. Text comparisons find 412 near-duplicate and 345 verbatim question pairs. Near-duplicates require question cosine similarity of at least 0.9 under BGE-large (Xiao et al., 2024),

confirmed with MPNet (Song et al., 2020). Table 11 lists the largest cross-benchmark duplicate flows, and Table 12 reports their directional shares.

Pool attacks. Attackers draw on a within-benchmark store of previously seen items or a cross-benchmark store preloaded with other benchmarks. Copying transfers the nearest stored answer and matches option text when choices differ. Retrieval supplies the eight nearest examples to Qwen3-8B (Yang et al., 2025a) at temperature 0. Online logistic regression predicts answer letters from BGE-large (Xiao et al., 2024) embeddings using five rotating held-out splits. Table 13 also summarizes a Laplace-smoothed position learner and two same-video sibling-item attackers. Full-scale runs use the reported item counts, and we track accuracy as evaluation proceeds.

Scoring and controls. We use 50 sequence permutations for copying and 3 for retrieval. Pool exploitability $\varepsilon_{\mathrm{pool}}$ is permutation-averaged accuracy over the final 20% of the sequence, smoothed over 20 items, minus chance. We obtain 95% intervals from 1,000 item-level bootstrap resamples. A control shuffles gold answers among items with equal option counts, preserving position priors while disrupting question-answer relations. Its excess accuracy is consistent with zero on 402 of 447 benchmark-attacker combinations, and fourteen retain a margin above 5%. The baseline c averages chance over the multiple-choice questions in each benchmark.

Reading the tables. Tables 14–15 list the target Benchmark and its full-scale item count n, rather than a 300-item sample. Superscript w denotes a store of previously seen items from the same benchmark, and superscript c a store preloaded with other benchmarks. The reported $\varepsilon_{\mathrm{pool}}$ is the largest final-stage margin among the listed attackers, in %. Att. identifies that attacker. L0 learns answer-position priors, L1 copies near-duplicates, L2-rule and L2-LLM use same-video sibling items through rules and a language model, and L3 retrieves eight examples. ALC in Table 14 is the selected attacker's exported normalized area under the learning curve, reported as a percentage. It summarizes the full sequence rather than only its final 20%. Zero denotes chance-level performance, and 100% denotes perfect accuracy throughout.

Diversity and robustness. Table 16 reports diversity, template entropy, and effective item counts. The table and Figure 11 cover 114 of the 115 audited benchmarks, since CG-Bench (Chen et al., 2025b) has no item embeddings. We compute Vendi scores on $n = 200$ items for a fixed-size comparison. The 8 smaller benchmarks use all available items and remain marked by their sample size. The primary embedding pairs SigLIP2 (Tschannen et al., 2025) with BGE-large (Xiao et al., 2024), and an alternative pairs DINOv3 (Siméoni et al., 2025) with MPNet (Song et al., 2020). Video embeddings average the same 32 frames, and text embeddings average token representations. Effective-size rankings agree across visual encoders at Spearman 0.88, across text encoders at 0.96, and across joint spaces at 0.90. Near-duplicate shares correlate at 0.90 across the two spaces. Diversity and judgment reliability capture different properties. Across 110 benchmarks, split-half reliability of the item-by-judge matrix correlates 0.04 with effective size and 0.00 with near-duplicate share.

Table 11: Largest cross-benchmark near-duplicate flows, confirmed in both embedding spaces.

Benchmark A	Benchmark B	Pairs	Exact
PAI-Bench-U	PhysicalAI	191	190
KFS-Bench	LongVideoBench	131	123
SITE-Bench	VSI-Bench	30	9
SciVideoBench	VideoKR	27	23
VideoKR	Video-MMMU	21	21
MMVU	VCR-Bench	16	14
MMVU	VideoKR	13	13
VKnowU	VSI-Bench	13	3
SpatialTree	VSI-Bench	10	7
Video-MME	VideoEval-Pro	9	9
VURB	Video-MME	7	7
MMVU	VURB	6	6
VURB	VideoKR	6	6
SITE-Bench	VURB	6	3

Table 11 (continued)

Benchmark A	Benchmark B	Pairs	Exact
GasVideo-1000	Video-MME	5	5
GasVideo-1000	MVBench	5	5
MVBench	TVBench	5	5
SpatialTree	VKnowU	5	3
SITE-Bench	SpatialTree	5	3
KFS-Bench	VideoEval-Pro	4	4
VCR-Bench	VideoKR	4	4
MVBench	VURB	4	4
VSI-Bench	VURB	4	1
SITE-Bench	Video-MME	4	4
LVBench	VideoEval-Pro	3	0
Perception Test	TVBench	3	1
TOMATO	TVBench	3	0
VCR-Bench	VURB	3	3
MVBench	SITE-Bench	2	2
MLVU	VideoEval-Pro	2	0

Table 12: Cross-benchmark near-duplicate shares (%). Rows supply items matched to columns. Column codes in row order are PA PhysicalAI, PU PAI-Bench-U, KF KFS-Bench, LV LongVideoBench, VK VideoKR, VS VSI-Bench, SI SITE-Bench, VU VURB, MM MMVU, SV SciVideoBench, VM Video-MME, VC VCR-Bench.

Items of	PA	PU	KF	LV	VK	VS	SI	VU	MM	SV	VM	VC
PhysicalAI	–	63.7	0	0	0	0	0	0	0	0	0	0
PAI-Bench-U	63.3	–	0	0	0	0	0	0	0	0	0	0
KFS-Bench	0	0	–	43.0	0	0	0	0	0	0	0	0
LongVideoBench	0	0	42.7	–	0	0	0	0	0	0	0	0
VideoKR	0	0	0	0	–	0	0	2.0	4.3	9.0	0	1.3
VSI-Bench	0	0	0	0	0	–	8.3	1.3	0	0	0	0
SITE-Bench	0	0	0	0	0	8.3	–	1.7	0	0	1.3	0
VURB	0	0	0	0	2.0	1.0	2.0	–	2.0	0	2.3	1.0
MMVU	0	0	0	0	4.3	0	0	2.0	–	0	0	5.3
SciVideoBench	0	0	0	0	8.7	0	0	0	0	–	0	0
Video-MME	0	0	0	0	0	0	1.3	2.3	0	0	–	0
VCR-Bench	0	0	0	0	1.3	0	0	0.3	5.3	0	0	–

Table 13: Pool attacks at full scale and on 300-item subsets, with learning-curve summaries.

Scale	Attacker	Pool	Median $\varepsilon_{\text{pool}}$ [IQR]	Median ALC	Expl.
full	L0 position prior	within	0.0 [-0.3, 1.9]	0.0	16/113
full	L1 near-duplicate copy	within	5.7 [1.5, 20.3]	7.1	46/113
full	L2-rule sibling-item rule	within	6.8 [2.0, 21.2]	7.8	46/113
full	L2-LLM sibling-item LLM	within	18.5 [8.1, 28.7]	17.1	72/113
full	L3 top-8 RAG	within	15.7 [7.9, 30.8]	20.2	80/113
full	any attacker (max)	within	23.7 [11.7, 37.6]	24.0	89/113
full	L0 position prior	cross	0.0 [-0.3, 0.9]	0.0	14/113
full	L1 near-duplicate copy	cross	9.3 [2.3, 28.1]	10.6	58/113
full	L2-rule sibling-item rule	cross	11.1 [2.7, 29.0]	11.5	60/113
full	L2-LLM sibling-item LLM	cross	20.0 [8.8, 33.8]	21.4	77/113
full	L3 top-8 RAG	cross	19.8 [7.8, 39.0]	25.0	82/113
full	any attacker (max)	cross	28.5 [13.5, 44.7]	30.0	91/113
s300	L0 position prior	within	0.1 [-0.5, 3.2]	0.0	8/113

Table 13 (continued)

Scale	Attacker	Pool	Median $\varepsilon_{\text{pool}}$ [IQR]	Median ALC	Expl.
s300	L1 near-duplicate copy	within	4.5 [0.9, 13.2]	4.0	25/113
s300	L2-rule sibling-item rule	within	5.4 [1.0, 14.9]	4.5	24/113
s300	L2-LLM sibling-item LLM	within	8.8 [3.6, 15.2]	6.5	34/113
s300	L3 top-8 RAG	within	14.9 [5.0, 24.9]	16.5	50/113
s300	any attacker (max)	within	15.4 [7.2, 27.6]	16.4	55/113
s300	L0 position prior	cross	0.0 [-0.9, 2.2]	0.0	8/113
s300	L1 near-duplicate copy	cross	6.6 [1.0, 13.1]	6.2	25/113
s300	L2-rule sibling-item rule	cross	6.7 [1.2, 15.7]	6.6	25/113
s300	L2-LLM sibling-item LLM	cross	8.8 [3.6, 16.3]	7.6	38/113
s300	L3 top-8 RAG	cross	14.0 [6.2, 25.9]	18.0	51/113
s300	any attacker (max)	cross	18.1 [8.7, 27.7]	19.0	56/113

Table 14: Within-pool and cross-pool exploitability across audited benchmarks (1/2).

Benchmark	n	$\varepsilon^{\text{w}}_{\text{pool}}$	Att.$^{\text{w}}$	ALC$^{\text{w}}$
TimeScope	3000	82.4	L2-LLM	92.1
LongTimeScope	900	81.2	L1	86.3
UCF101-AD	4224	76.0	L3	81.6
Video-TT	5000	73.2	L2-LLM	78.8
STI-Bench	4128	71.1	L2-rule	54.2
FineBench	20143	69.2	L1	90.5
SAW-Bench	4142	67.7	L2-rule	70.9
VF-Eval	5932	67.3	L2-rule	50.0
VURB	2100	64.3	L1	71.2
VAU-Bench	251	64.2	L3	88.1
VRU-Accident	5989	56.3	L1	72.4
VideoMathQA	2100	55.1	L2-rule	79.3
VANE-Bench	559	53.3	L2-LLM	47.8
GUI-World	3642	46.9	L3	60.6
Neptune	5000	45.9	L1	40.4
TaskMeAnything-v1	3567	45.8	L2-rule	47.7
MMWorld	1559	42.1	L3	52.5
ForeSeaQA	591	41.7	L2-LLM	48.9
ALLVB	25564	40.8	L3	49.0
Video-Holmes	3388	40.4	L2-LLM	36.1
VELOCITI	8789	39.0	L2-rule	50.6
HLV-1K	10533	38.9	L3	50.5
SLVMBench	2255	38.5	L3	44.7
LongVideo-Reason	856	37.4	L3	52.9
VGenST-Bench	9707	35.1	L3	46.9
FAVOR-Bench	5000	34.8	L2-LLM	29.9
MMR-V	1256	34.8	L2-LLM	25.8
CMU	4153	33.5	L2-LLM	22.8
ViMUL-Bench	5000	31.0	L2-LLM	35.0
Colon-Bench	1485	30.8	L3	34.3
ExAct	3521	29.3	L3	33.7
ReXTime	921	28.7	L3	26.4
StreamingBench	2500	28.5	L2-LLM	31.4
DSI-Bench	1769	27.9	L1	35.5
MLVU	2174	27.0	L3	31.1
Causal2Needles	747	26.3	L3	30.1
Q-Bench-Video	892	25.9	L3	49.0
HERBench	1971	25.9	L2-LLM	30.5
SFI-Bench	1555	25.7	L3	30.2
Video-MMMU	900	25.6	L2-LLM	22.1
COVER	5000	25.5	L3	33.0

Table 14 (continued)

Benchmark	n	$\varepsilon^{\mathrm{w}}_{\mathrm{pool}}$	Att.$^{\mathrm{w}}$	ALC$^{\mathrm{w}}$
CAViAR	2996	25.2	L3	38.6
SIS-Bench	4856	23.6	L2-LLM	20.8
SpatialTree	814	23.0	L3	21.8
UrbanVideo-Bench	5355	23.0	L3	26.5
SciVideoBench	1000	22.6	L2-LLM	18.2
VideoEspresso	1382	20.1	L2-LLM	21.7
VCR-Bench	2068	19.6	L1	24.6
EgoSchema	500	18.9	L3	24.8
M3-Verse	3374	18.7	L3	21.3
K9-Bench	4744	18.5	L2-LLM	16.4
Video-MME	2700	18.5	L2-LLM	17.2
MotionBench	4018	17.5	L3	42.5
Video-LevelGauge	1177	17.3	L2-LLM	13.5
LongVideoBench	815	17.0	L3	23.5
Perception Test	5000	15.9	L2-rule	21.8
ReVSI	5000	15.8	L1	24.0
MMVU	1000	15.8	L3	23.6
VideoAds	1100	15.4	L2-LLM	9.1
OVO-S-Bench	1632	15.0	L3	14.0
Artifact-Bench	850	14.1	L1	-8.1
CineTechBench	127	13.5	L3	9.5
SmartHome-Bench	1023	13.2	L0	25.3
VITATECS	5000	13.1	L3	20.5
iSafetyBench	3456	12.8	L3	12.1
TVBench	2525	12.6	L1	14.7
KFS-Bench	580	12.2	L2-LLM	7.7
SITE-Bench	3619	12.1	L2-rule	12.1
PhysicalAI	1214	11.8	L3	36.8
ScaleLong	1723	11.7	L2-LLM	12.9
PAI-Bench-U	1214	11.7	L3	36.8
Video-MME-v2	2865	11.7	L2-LLM	8.2
VKnowU	1680	11.2	L3	10.9
VSI-Bench	5000	10.8	L1	20.7
TOMATO	1484	10.8	L2-LLM	10.3
MME-VideoOCR	1456	10.2	L3	11.0
TemporalBench	5000	10.2	L2-LLM	48.3
MVBench	4000	9.1	L2-LLM	9.7
VideoOdyssey	1618	8.4	L3	9.0
ContPhy	1715	7.9	L1	0.5
GasVideo-1000	1013	7.6	L2-rule	19.0
VideoEval-Pro	1289	7.1	L2-LLM	5.0
DSR-Bench	1450	7.1	L2-rule	8.6
ST-Evidence	4004	6.6	L3	24.2
SoMi-ToM	75	6.4	L3	43.8
ExpVid	7800	6.4	L3	15.4
ReactHuman	1054	5.7	L0	7.6
TOC-Bench	2321	5.6	L3	7.9
V2P-Bench	1169	5.5	L3	31.7
VSI-Super-Wild	5000	4.5	L2-LLM	7.8
VideoKR	2000	4.5	L3	8.3
LVBench	1549	4.4	L2-LLM	27.7
v-HUB	1902	3.7	L2-LLM	15.9
MMSI-Video-Bench	1106	3.6	L3	6.1
E.T. Bench	1500	3.6	L3	3.6
EgoPlan-Bench	3343	3.5	L2-LLM	2.4
S-EMBER	9448	3.2	L3	4.3
VNBench	320	3.1	L3	0.8
HourVideo	1157	1.8	L3	1.5
IntPhys 2	1072	1.1	L2-rule	0.4

Table 14 (continued)

Benchmark	n	$\varepsilon^{\mathrm{w}}_{\mathrm{pool}}$	Att.$^{\mathrm{w}}$	ALC$^{\mathrm{w}}$
VSTAT	1500	0.6	L1	16.2
PLM-VideoBench	665	0.0	L0	0.0
LEMONADE	5000	-0.1	L2-rule	24.9
Vinoground	1000	-3.8	L0	-3.3
VidPair-Halluc	5000	-3.9	L1	18.2
TempCompass	5000	-5.3	L1	17.1
AoTBench	5000	-6.7	L3	43.3
MORSE-500	500	-10.2	L3	8.1
OSI-Bench	5000	-12.2	L1	5.1
NaturalBench-Video	1860	-15.8	L2-LLM	29.8
PerceptionComp	1114	-16.6	L2-LLM	2.4
TimeBlind	2400	-21.3	L2-LLM	24.5
VideoSIAH-Eval	652	-24.8	L2-LLM	0.1

Table 15: Within-pool and cross-pool exploitability across audited benchmarks (2/2).

Benchmark	n	$\varepsilon^{\mathrm{c}}_{\mathrm{pool}}$	Att.$^{\mathrm{c}}$
TimeScope	3000	82.5	L2-LLM
LongTimeScope	900	82.3	L2-LLM
UCF101-AD	4224	76.0	L3
Video-TT	5000	73.2	L2-LLM
STI-Bench	4128	70.9	L2-rule
FineBench	20143	69.2	L1
SAW-Bench	4142	67.7	L2-rule
VF-Eval	5932	67.5	L2-rule
VURB	2100	69.8	L1
VAU-Bench	251	63.7	L3
VRU-Accident	5989	56.6	L1
VideoMathQA	2100	55.2	L2-rule
VANE-Bench	559	53.3	L2-LLM
GUI-World	3642	46.8	L2-LLM
Neptune	5000	45.9	L1
TaskMeAnything-v1	3567	45.8	L2-rule
MMWorld	1559	40.8	L3
ForeSeaQA	591	41.7	L2-LLM
ALLVB	25564	40.3	L3
Video-Holmes	3388	40.4	L2-LLM
VELOCITI	8789	39.0	L2-rule
HLV-1K	10533	37.9	L3
SLVMBench	2255	39.0	L3
LongVideo-Reason	856	34.9	L3
VGenST-Bench	9707	34.8	L3
FAVOR-Bench	5000	34.8	L2-LLM
MMR-V	1256	34.2	L2-LLM
CMU	4153	34.7	L2-LLM
ViMUL-Bench	5000	31.0	L2-LLM
Colon-Bench	1485	31.1	L3
ExAct	3521	29.3	L3
ReXTime	921	44.7	L3
StreamingBench	2500	28.5	L2-LLM
DSI-Bench	1769	28.1	L1
MLVU	2174	34.2	L3
Causal2Needles	747	26.6	L3
Q-Bench-Video	892	24.3	L3
HERBench	1971	25.9	L2-LLM
SFI-Bench	1555	24.4	L2-LLM

Table 15 (continued)

Benchmark	n	$\varepsilon^{c}_{\text{pool}}$	Att.[c]
Video-MMMU	900	45.5	L3
COVER	5000	24.3	L3
CAViAR	2996	25.2	L3
SIS-Bench	4856	23.3	L2-LLM
SpatialTree	814	33.0	L3
UrbanVideo-Bench	5355	22.9	L3
SciVideoBench	1000	58.6	L3
VideoEspresso	1382	20.1	L2-LLM
VCR-Bench	2068	19.6	L1
EgoSchema	500	19.8	L3
M3-Verse	3374	18.7	L3
K9-Bench	4744	18.5	L2-LLM
Video-MME	2700	31.1	L3
MotionBench	4018	20.0	L3
Video-LevelGauge	1177	18.4	L3
LongVideoBench	815	18.5	L3
Perception Test	5000	16.8	L1
ReVSI	5000	15.8	L1
MMVU	1000	30.3	L1
VideoAds	1100	15.4	L2-LLM
OVO-S-Bench	1632	16.5	L3
Artifact-Bench	850	13.7	L1
CineTechBench	127	13.5	L3
SmartHome-Bench	1023	13.2	L0
VITATECS	5000	13.1	L3
iSafetyBench	3456	12.8	L3
TVBench	2525	19.6	L3
KFS-Bench	580	47.5	L1
SITE-Bench	3619	34.8	L1
PhysicalAI	1214	57.2	L1
ScaleLong	1723	11.7	L2-LLM
PAI-Bench-U	1214	57.0	L1
Video-MME-v2	2865	11.8	L2-LLM
VKnowU	1680	16.6	L3
VSI-Bench	5000	16.9	L1
TOMATO	1484	11.1	L2-LLM
MME-VideoOCR	1456	10.1	L3
TemporalBench	5000	10.7	L3
MVBench	4000	22.3	L2-LLM
VideoOdyssey	1618	8.1	L2-LLM
ContPhy	1715	7.7	L1
GasVideo-1000	1013	24.8	L2-rule
VideoEval-Pro	1289	68.4	L3
DSR-Bench	1450	8.9	L2-rule
ST-Evidence	4004	7.2	L3
SoMi-ToM	75	6.4	L3
ExpVid	7800	6.6	L3
ReactHuman	1054	5.7	L1
TOC-Bench	2321	5.7	L3
V2P-Bench	1169	32.4	L3
VSI-Super-Wild	5000	4.5	L2-LLM
VideoKR	2000	29.7	L2-rule
LVBench	1549	38.2	L3
v-HUB	1902	3.7	L2-LLM
MMSI-Video-Bench	1106	1.8	L2-LLM
E.T. Bench	1500	3.6	L3
EgoPlan-Bench	3343	3.5	L2-LLM
S-EMBER	9448	3.0	L3
VNBench	320	3.6	L1

Table 15 (continued)

Benchmark	n	$\varepsilon^{c}_{\text{pool}}$	Att.[c]
HourVideo	1157	1.6	L1
IntPhys 2	1072	1.1	L2-rule
VSTAT	1500	0.6	L1
PLM-VideoBench	665	0.0	L0
LEMONADE	5000	-0.1	L2-rule
Vinoground	1000	-0.8	L0
VidPair-Halluc	5000	-3.8	L1
TempCompass	5000	-5.3	L1
AoTBench	5000	-2.1	L3
MORSE-500	500	-10.2	L3
OSI-Bench	5000	-12.2	L1
NaturalBench-Video	1860	-15.8	L2-LLM
PerceptionComp	1114	-16.6	L2-LLM
TimeBlind	2400	-21.3	L2-LLM
VideoSIAH-Eval	652	-24.8	L2-LLM

Table 16: Content diversity, duplication, template entropy, and effective item counts by benchmark.

Benchmark	VS/n	NearDup %	SameVid %	TemplEnt	AnsPosEnt	EffN
ALLVB	0.170	24.3	0.7	1.00	0.83	34
AoTBench	0.161	35.0	0.0	0.95	–	32
Artifact-Bench	0.112	49.3	0.0	0.13	1.43	22
Causal2Needles	0.137	41.7	0.8	1.00	1.00	27
CAViAR	0.022	100.0	0.1	0.19	1.22	4
CineTechBench	0.067	87.6	0.0	0.00	0.99	8
CMU	0.163	28.3	0.1	0.59	1.66	33
Colon-Bench	0.042	99.8	0.2	1.00	1.00	8
ContPhy	0.046	100.0	0.1	0.94	1.56	9
COVER	0.258	7.8	0.0	0.95	0.85	52
DSI-Bench	0.092	67.3	0.1	0.25	1.00	18
DSR-Bench	0.155	49.4	0.1	0.84	0.99	31
E.T. Bench	0.062	97.1	0.0	0.68	1.11	12
EgoPlan-Bench	0.043	97.0	0.0	0.96	–	9
EgoSchema	0.119	66.0	0.0	1.00	0.99	24
ExAct	0.022	100.0	0.0	0.00	0.99	4
ExpVid	0.109	69.1	0.0	0.85	1.09	22
FAVOR-Bench	0.195	25.7	0.1	1.00	0.99	39
FineBench	0.039	100.0	25.5	0.91	0.99	6
ForeSeaQA	0.134	84.3	4.6	0.97	0.96	27
GasVideo-1000	0.272	39.2	0.4	0.92	1.18	54
GUI-World	0.268	11.6	0.0	1.00	0.79	54
HERBench	0.059	99.2	1.8	0.81	0.98	12
HLV-1K	0.164	46.1	0.1	0.94	0.79	33
HourVideo	0.141	73.8	2.2	0.87	0.99	28
IntPhys 2	0.024	100.0	0.0	0.00	1.00	5
iSafetyBench	0.050	98.6	0.1	0.12	0.98	10
K9-Bench	0.146	41.0	0.1	1.00	1.00	29
KFS-Bench	0.255	26.8	0.3	0.98	0.99	51
LEMONADE	0.047	99.9	2.1	0.72	0.99	9
LongTimeScope	0.054	100.0	0.1	0.67	–	11
LongVideo-Reason	0.214	10.7	0.0	1.00	0.83	43
LongVideoBench	0.303	15.3	0.1	0.96	0.99	55
LVBench	0.152	67.5	0.9	1.00	–	30
M3-Verse	0.019	100.0	0.4	0.95	1.16	4
MLVU	0.108	76.7	0.1	0.98	0.99	22
MME-VideoOCR	0.347	9.5	0.0	0.99	0.99	70

Table 16 (continued)

Benchmark	VS/n	NearDup %	SameVid %	TemplEnt	AnsPosEnt	EffN
MMR-V	0.202	24.2	0.4	1.00	1.00	40
MMSI-Video-Bench	0.158	47.9	0.1	1.00	1.09	32
MMVU	0.247	25.4	0.2	1.00	0.99	50
MMWorld	0.234	52.6	0.9	0.94	0.94	47
MORSE-500	0.025	100.0	0.0	0.00	–	5
MotionBench	0.222	29.4	0.0	0.96	–	44
MVBench	0.152	67.2	0.0	0.86	0.97	30
NaturalBench-Video	0.217	67.3	0.2	0.98	–	44
Neptune	0.279	13.9	0.1	0.97	–	56
NExT-QA	0.262	7.1	0.1	0.99	–	52
OSI-Bench	0.083	91.2	0.1	0.79	0.97	16
OVO-S-Bench	0.082	91.3	1.0	0.93	0.99	16
PAI-Bench-U	0.221	46.4	0.0	0.96	1.46	44
Perception Test	0.160	77.4	0.0	0.76	1.00	32
PerceptionComp	0.168	66.6	0.5	1.00	–	34
PhysicalAI	0.221	45.1	0.0	0.95	–	44
PLM-VideoBench	0.242	19.3	0.0	0.95	–	48
Q-Bench-Video	0.192	12.1	0.0	0.99	–	38
ReactHuman	0.013	100.0	0.0	0.00	1.00	2
ReVSI	0.095	93.7	0.1	0.67	0.97	19
ReXTime	0.309	6.2	0.1	1.00	1.00	62
S-EMBER	0.216	12.0	0.0	1.00	0.99	43
SAW-Bench	0.040	97.3	0.2	0.34	0.99	8
ScaleLong	0.266	28.4	0.3	1.00	1.00	53
SciVideoBench	0.082	76.1	0.4	0.98	0.99	16
SFI-Bench	0.140	59.0	0.9	1.00	1.00	28
SIS-Bench	0.056	84.9	0.1	0.56	1.00	11
SITE-Bench	0.205	46.8	0.1	0.88	1.05	41
SLVMBench	0.201	51.3	0.0	0.99	0.99	40
SmartHome-Bench	0.022	98.6	0.0	0.00	0.95	4
SoMi-ToM	0.047	100.0	6.8	0.87	–	3
SpatialTree	0.081	78.6	0.0	0.33	0.95	16
ST-Evidence	0.242	34.9	0.0	0.96	1.24	48
STI-Bench	0.095	87.5	0.3	0.83	0.98	19
StreamingBench	0.205	54.2	0.1	0.93	0.99	41
TaskMeAnything-v1	0.080	84.4	0.0	0.74	0.99	16
TempCompass	0.172	58.3	0.3	1.00	–	34
TemporalBench	0.144	62.3	0.0	0.98	–	29
TimeBlind	0.187	80.2	0.3	0.98	–	33
TimeScope	0.066	99.8	0.0	0.69	–	13
TOC-Bench	0.164	41.7	0.0	0.93	0.95	33
TOMATO	0.147	71.6	0.0	0.83	0.97	29
TVBench	0.093	90.3	0.0	0.61	0.95	18
UCF101-AD	0.064	92.1	0.0	0.00	0.99	13
UrbanVideo-Bench	0.067	94.3	0.1	0.98	1.08	13
v-HUB	0.211	22.0	0.1	1.00	–	42
V2P-Bench	0.194	25.1	0.1	1.00	–	39
VANE-Bench	0.156	72.1	0.2	0.82	0.80	31
VAU-Bench	0.116	69.8	0.0	0.98	0.66	23
VCR-Bench	0.339	9.0	0.1	1.00	1.07	53
VELOCITI	0.021	99.1	0.1	0.00	1.00	4
VF-Eval	0.184	30.1	0.1	0.80	1.00	37
VGenST-Bench	0.252	28.2	0.1	0.81	0.94	50
Video-Holmes	0.159	36.3	0.2	0.96	0.99	32
Video-LevelGauge	0.341	22.1	0.2	1.00	0.99	68
Video-MME	0.267	5.3	0.1	0.99	0.99	53
Video-MME-v2	0.309	15.0	0.1	1.00	0.99	62
Video-MMMU	0.299	22.9	0.2	1.00	0.94	60
Video-TT	0.308	12.7	0.1	1.00	–	62

Table 16 (continued)

Benchmark	VS/n	NearDup %	SameVid %	TemplEnt	AnsPosEnt	EffN
VideoAds	0.200	27.4	0.5	0.98	1.00	40
VideoEspresso	0.262	23.3	0.0	1.00	0.89	52
VideoEval-Pro	0.296	16.9	0.5	1.00	1.04	59
VideoKR	0.249	25.6	0.1	1.00	0.93	50
VideoMathQA	0.200	63.3	0.2	0.95	1.54	40
VideoOdyssey	0.303	27.1	1.4	1.00	0.99	42
VideoSIAH-Eval	0.319	8.1	0.3	1.00	–	64
VidPair-Halluc	0.135	75.0	0.1	0.80	0.99	27
ViMUL-Bench	0.284	9.6	0.1	0.93	0.99	57
Vinoground	0.060	81.4	0.0	0.00	1.00	12
VITATECS	0.062	78.8	0.0	0.00	1.00	12
VKnowU	0.227	41.7	0.1	0.80	1.60	45
VNBench	0.125	57.7	0.0	0.38	1.00	25
VRU-Accident	0.023	99.9	0.1	0.34	1.00	5
VSI-Bench	0.075	97.8	0.6	0.49	0.99	15
VSI-Super-Wild	0.044	97.0	0.3	0.33	0.99	9
VSTAT	0.062	100.0	0.1	0.70	0.99	12
VURB	0.219	60.9	0.9	0.93	1.03	44

L SINGLE FRAMES AND CAPTIONS

Frame-level protocol. The frame-level attackers replace video with a single image or independently generated frame captions. Gemini 3.5 Flash (Google DeepMind, 2026b) reads the normalized video's middle frame, and Claude Opus 5 (Anthropic, 2026b) reads the captions. We compare each attack with the reference on their commonly scored items and subtract chance accuracy. The larger margin defines $\varepsilon_{\text{frame}}$. This tests whether a benchmark requires information beyond one image or a sequence of frame descriptions.

Caption construction. AuroraCap-7B-IMG (Chai et al., 2025) describes each of the 32 sampled frames independently. Each call receives one frame and a fixed instruction to describe the visible objects, people, actions, and setting in one short sentence without guessing what happened before or after. The captioner never receives the question, the options, or the other frames. The reader receives these descriptions in frame order, without timestamps, followed by the question and options. Its instruction begins `You are given a text description of a video INSTEAD of the video itself.` and requests an answer letter. Table 7 separately reports a whole-video caption generated by Qwen3-VL-8B (Bai et al., 2025) with the prompt `Describe this video in detail based on the frames provided, in temporal order.` This whole-video description enters no attack level, and the caption attack uses only the independently generated frame descriptions.

Diagnostic conditions. Tables 17 and 18 provide a broader diagnostic sweep with Qwen3-VL-8B. In Table 17, blind supplies no video, 1f uses position 16 of the 32-frame grid, and 32f and 128f give frame counts. The 448px and 224px columns vary resolution, shuf permutes frame order, half1 and half2 restrict input to each video half, and off1 and off2 shift the sampling grid. Table 18 summarizes the derived capability profiles. These conditions help interpret sensitivity to visual detail and temporal coverage, and not all of them are frame-level attacks. In particular, the 224-pixel half-resolution probe remains outside this level. Dashes mark unavailable measurements or derived scores.

What the comparisons show. As shown in Figure 4, frame captions recover more reference accuracy than a single image. Because the descriptions cover multiple moments in order, this result tests substitution by text with the temporal information intact. The diagnostic sweep also shows that single-frame sufficiency varies across benchmarks. On MMWorld (He et al., 2024), accuracy is 67% with one frame and 68% with 32. On EgoSchema (Mangalam et al., 2023), it rises from 59% to 78%. These contrasts motivate separate checks for static visual cues, information spread across frames, and dependence on frame order.

Table 17: Qwen3-VL-8B condition accuracy (%), with blank cells for probes outside a benchmark's sweep.

Benchmark	blind	1f	32f	448px	224px	128f	shuf	half1	half2	off1	off2
AoTBench	49	48	65	65	61	66	51	60	58	63	64
Artifact-Bench	0	2	1	1	1	–	1	–	–	–	–
COVER	45	51	59	59	56	52	54	50	46	56	56
DSI-Bench	47	49	59	60	55	75	0	50	50	75	75
DSR-Bench	32	32	32	32	32	34	33	34	33	32	33
ST-Evidence	57	64	79	79	77	69	70	58	69	68	71
EgoSchema	36	59	78	77	75	78	76	70	74	74	75
ExAct	38	41	42	42	41	30	33	34	30	33	33
ExpVid	33	43	46	48	44	43	42	46	40	44	42
FAVOR-Bench	40	35	53	51	52	51	41	40	44	51	52
GasVideo-1000	38	51	69	66	62	70	66	62	60	70	70
HourVideo	20	20	20	21	20	21	19	18	19	20	20
IntPhys 2	51	47	48	47	50	–	–	–	–	–	–
LVBench	21	24	37	36	38	48	39	33	36	38	38
LEMONADE	31	29	34	34	34	32	35	34	33	33	33
LongTimeScope	34	26	27	27	27	30	26	24	26	25	25
VideoSIAH-Eval	3	2	10	10	8	14	9	8	5	9	9
LongVideoBench	37	39	64	64	66	63	60	56	53	59	62
MLVU	37	40	70	68	66	76	65	54	58	66	65
MMR-V	38	39	44	45	43	51	46	43	42	46	46
MMSI-Video-Bench	25	26	27	28	29	28	29	29	28	28	28
MMVU	49	56	71	70	70	68	71	71	62	72	70
MMWorld	53	67	68	69	68	75	75	72	72	75	76
MORSE-500	0	3	27	28	31	23	21	3	26	20	25
MVBench	31	48	66	66	65	71	66	55	61	72	71
MotionBench	40	51	65	64	61	51	47	42	51	51	51
NExT-QA	46	72	81	82	80	85	84	79	80	84	83
Neptune	6	6	24	18	12	18	24	19	18	18	24
OSI-Bench	36	35	41	46	44	42	41	44	39	44	44
PAI-Bench-U	26	41	43	48	45	50	41	50	50	50	50
PerceptionComp	3	4	9	10	10	12	8	9	7	9	11
Perception Test	37	54	70	70	65	72	68	65	65	71	73
Q-Bench-Video	53	46	53	54	54	0	45	50	0	50	0
ReVSI	30	37	48	46	44	48	51	45	39	48	46
SAW-Bench	33	23	31	31	29	30	33	30	25	29	31
SFI-Bench	49	32	57	58	55	59	55	50	49	60	56
SIS-Bench	43	54	65	64	63	54	46	42	45	56	56
STI-Bench	38	38	43	43	38	42	40	42	42	43	42
SciVideoBench	18	21	27	27	26	29	27	27	25	29	27
SITE-Bench	27	37	63	62	61	61	60	55	55	63	64
SoMi-ToM	47	52	87	83	57	83	83	62	83	88	88
SpatialTree	37	30	38	38	40	31	28	31	34	28	31
StreamingBench	48	48	66	66	63	71	66	63	61	68	67
TOMATO	23	22	31	29	30	8	25	17	33	17	25
TVBench	37	37	50	51	48	53	37	46	40	50	52
TempCompass	6	16	24	26	26	38	42	33	42	38	38
TemporalBench	55	56	64	65	64	56	60	62	63	63	61
TimeBlind	39	37	58	55	55	55	36	27	64	55	55
TimeScope	40	33	58	59	61	73	57	45	39	59	58
V2P-Bench	34	34	51	52	51	60	48	50	45	54	53
VCR-Bench	38	36	52	53	50	54	48	47	46	53	50
VITATECS	64	79	87	87	86	89	89	86	86	89	89
VKnowU	41	45	66	67	66	66	65	62	64	67	65
VSI-Bench	40	34	60	60	60	64	51	55	47	64	56
VSI-Super-Wild	31	28	33	33	33	33	31	34	28	32	32
VSTAT	38	37	41	39	39	40	40	40	42	40	41

Table 17 (continued)

Benchmark	blind	1f	32f	448px	224px	128f	shuf	half1	half2	off1	off2
VURB	38	29	50	52	52	55	55	43	42	51	50
ViMUL-Bench	31	46	63	62	64	66	59	54	60	60	61
VidPair-Halluc	27	44	47	47	46	–	43	–	–	–	–
Video-Holmes	38	35	49	49	49	54	49	43	46	49	51
Video-MME	40	42	66	66	63	69	64	59	55	65	64
Video-MME-v2	17	19	28	27	24	32	26	28	25	29	27
Video-TT	13	15	8	13	12	5	14	12	14	2	9
VideoEval-Pro	26	27	42	41	43	52	43	37	39	41	44
VideoKR	19	23	36	37	32	39	35	30	33	33	32
Video-MMMU	38	43	58	56	54	58	54	44	53	53	54
VideoMathQA	48	50	55	54	56	56	57	53	54	55	56
VideoOdyssey	32	26	36	36	32	41	36	32	31	37	40
Vinoground	49	53	73	73	71	71	29	57	29	71	71
v-HUB	20	49	59	58	58	52	48	48	41	52	52
ALLVB	59	56	62	–	62	67	61	63	64	65	62
Causal2Needles	43	39	52	–	51	66	53	54	45	54	56
CAViAR	26	73	71	–	69	69	70	72	71	72	72
CG-Bench	20	26	43	–	38	–	42	–	–	39	40
CineTechBench	32	53	50	–	50	–	43	–	–	–	–
Colon-Bench	27	40	51	–	47	48	48	48	49	48	50
ContPhy	18	21	30	–	32	–	28	–	–	–	–
E.T. Bench	22	32	53	–	48	43	32	42	34	54	54
EgoPlan-Bench	26	26	27	–	24	–	24	–	–	27	26
FineBench	58	45	54	–	56	58	55	55	56	54	54
ForeSeaQA	63	58	60	–	59	66	61	56	58	64	62
GUI-World	74	78	89	–	87	86	85	75	84	88	87
HERBench	47	50	58	–	58	55	54	52	55	54	56
HLV-1K	60	46	64	–	68	75	67	67	55	66	66
iSafetyBench	16	33	39	–	39	–	40	–	–	–	–
K9-Bench	37	39	51	–	53	57	52	50	48	54	54
KFS-Bench	39	40	58	–	55	56	53	48	46	54	56
LongVideo-Reason	69	65	74	–	74	77	76	70	71	75	74
M3-Verse	34	35	39	–	35	41	39	34	38	37	41
MME-VideoOCR	25	49	65	–	55	77	67	66	59	67	71
NaturalBench-Video	51	55	71	–	71	83	57	83	67	83	83
OVO-S-Bench	35	31	42	–	42	44	40	38	34	41	40
PLM-VideoBench	59	71	74	–	70	–	67	–	–	67	64
ReactHuman	31	34	47	–	46	78	53	72	39	56	56
ReXTime	42	66	82	–	82	80	81	77	80	82	80
S-EMBER	16	11	26	–	21	34	27	21	21	24	28
ScaleLong	36	40	52	–	52	57	51	54	44	56	50
SLVMBench	38	43	63	–	55	68	59	60	50	64	64
SmartHome-Bench	37	64	70	–	70	71	68	73	68	69	69
TaskMeAnything-v1	26	37	50	–	48	62	66	59	64	65	68
TOC-Bench	22	22	25	–	26	26	24	24	29	27	26
UCF101-AD	84	50	47	–	46	–	48	–	–	–	–
UrbanVideo-Bench	32	37	41	–	39	41	41	34	41	41	42
VANE-Bench	72	72	76	–	77	–	76	–	–	–	–
VAU-Bench	95	94	98	–	97	96	96	97	96	96	97
VELOCITI	61	60	86	–	84	–	82	–	–	–	–
VF-Eval	46	53	59	–	59	–	58	–	–	–	–
VGenST-Bench	41	50	72	–	70	–	65	–	–	–	–
Video-LevelGauge	30	53	89	–	83	89	89	76	78	90	90
VideoAds	27	33	64	–	64	68	65	54	62	65	64
VideoEspresso	37	51	59	–	60	58	61	58	57	56	59
VNBench	25	23	62	–	63	79	41	40	24	57	49
VRU-Accident	50	60	68	–	66	78	73	73	62	71	73

Table 18: Qwen3-VL-8B diagnostic capability profiles across audited benchmarks

Benchmark	m	S	T_{coarse}	O	L	PosBias	Φ_{flip}	PriorDom
AoTBench	0.166	0.245	0.089	0.801	0.985	0.020	0.069	no
Artifact-Bench	0.008	–	–	–	–	–	–	yes
COVER	0.140	0.219	-0.413	0.275	0.967	0.038	0.019	no
DSI-Bench	0.117	0.314	-0.000	6.407	0.667	0.000	0.000	no
DSR-Bench	-0.007	–	–	–	0.979	0.013	0.053	yes
ST-Evidence	0.221	-0.024	-0.073	0.000	0.966	-0.113	0.057	no
EgoSchema	0.420	0.071	-0.000	0.032	0.987	-0.033	0.117	no
ExAct	0.040	–	–	–	1.000	0.038	0.048	yes
ExpVid	0.129	-0.244	0.108	0.000	0.967	0.056	0.028	no
FAVOR-Bench	0.130	0.146	0.095	0.665	0.963	-0.043	0.086	no
GasVideo-1000	0.306	0.117	-0.012	0.143	0.953	0.015	0.080	no
HourVideo	0.003	–	–	–	0.787	-0.010	0.140	yes
IntPhys 2	-0.037	–	–	–	–	–	–	yes
LVBench	0.167	-0.020	0.620	-0.080	0.812	-0.023	0.370	no
LEMONADE	0.030	–	–	–	0.932	0.013	0.150	yes
LongTimeScope	-0.077	–	–	–	0.825	-0.013	0.127	yes
VideoSIAH-Eval	0.065	-0.078	0.667	0.133	0.636	0.030	0.104	no
LongVideoBench	0.273	0.021	0.071	0.024	0.925	0.033	0.221	no
MLVU	0.330	0.057	0.182	0.131	0.871	-0.047	0.233	no
MMR-V	0.067	-0.052	0.684	0.105	0.916	0.007	0.131	no
MMSI-Video-Bench	0.023	–	–	–	0.955	0.008	0.114	yes
MMVU	0.218	0.049	-0.168	0.056	0.949	0.085	0.043	no
MMWorld	0.151	-0.050	-0.000	-0.041	0.967	-0.006	0.055	no
MORSE-500	0.269	-0.138	0.122	-0.061	0.917	-0.230	0.115	no
MVBench	0.350	0.038	-0.000	0.157	0.942	-0.062	0.055	no
MotionBench	0.250	0.027	-0.000	0.151	1.000	-0.094	0.038	no
NExT-QA	0.353	-0.019	0.055	-0.033	0.986	-0.012	0.062	no
Neptune	0.176	0.354	-0.333	0.000	1.000	0.062	0.059	no
OSI-Bench	0.044	–	–	–	0.909	0.050	0.037	yes
PAI-Bench-U	0.174	-0.311	-0.000	0.111	1.000	0.000	0.000	no
PerceptionComp	0.064	-0.333	0.526	0.105	0.778	0.017	0.087	no
Perception Test	0.327	-0.020	0.050	0.076	0.942	0.000	0.062	no
Q-Bench-Video	0.000	–	–	–	1.000	0.500	0.500	yes
ReVSI	0.180	0.096	-0.000	-0.182	0.845	0.057	0.172	no
SAW-Bench	-0.015	–	–	–	0.859	0.048	0.051	yes
SFI-Bench	0.087	-0.077	0.192	0.231	0.930	0.007	0.217	no
SIS-Bench	0.217	0.031	-0.044	0.440	0.948	-0.029	0.009	no
STI-Bench	0.043	–	–	–	0.945	0.000	0.054	yes
SciVideoBench	0.090	0.049	0.222	-0.037	0.864	0.017	0.187	no
SITE-Bench	0.354	0.033	0.013	0.038	0.925	-0.004	0.196	no
SoMi-ToM	0.383	0.783	-0.087	0.087	1.000	-0.217	0.133	no
SpatialTree	0.016	–	–	–	0.889	-0.035	0.190	yes
StreamingBench	0.183	0.020	0.260	0.040	0.950	0.026	0.187	no
TOMATO	0.077	0.301	-0.000	-2.167	1.000	-0.167	0.167	no
TVBench	0.133	-0.049	0.362	0.797	0.899	0.063	0.111	no
TempCompass	0.179	-0.143	-0.000	-0.232	1.000	-0.083	0.000	no
TemporalBench	0.088	-0.239	-0.131	-0.261	0.980	-0.011	0.081	no
TimeBlind	0.227	0.121	-0.000	0.800	1.000	-0.364	0.000	no
TimeScope	0.183	-0.051	0.800	0.054	0.840	0.063	0.193	no
V2P-Bench	0.167	-0.286	0.528	0.192	0.915	0.044	0.228	no
VCR-Bench	0.138	-0.083	0.128	0.255	0.912	0.007	0.165	no
VITATECS	0.237	0.042	-0.000	0.000	1.000	0.000	0.000	no
VKnowU	0.246	-0.118	0.038	0.000	0.971	-0.019	0.100	no
VSI-Bench	0.199	0.035	0.172	0.483	0.886	0.082	0.294	no
VSI-Super-Wild	0.024	–	–	–	0.868	0.059	0.194	yes
VSTAT	0.028	–	–	–	0.923	-0.023	0.061	yes
VURB	0.112	-0.507	0.426	-0.480	0.845	0.006	0.327	no

Table 18 (continued)

Benchmark	m	S	T_{coarse}	O	L	PosBias	Φ_{flip}	PriorDom
ViMUL-Bench	0.324	-0.018	0.153	0.066	0.919	-0.057	0.142	no
VidPair-Halluc	0.202	0.000	–	0.191	–	–	–	no
Video-Holmes	0.113	0.000	0.471	0.000	0.925	-0.030	0.213	no
Video-MME	0.263	0.023	0.115	0.077	0.918	0.040	0.219	no
Video-MME-v2	0.105	0.069	0.414	0.172	0.870	0.029	0.174	no
Video-TT	-0.050	–	–	–	0.500	-0.023	0.093	yes
VideoEval-Pro	0.163	-0.020	0.592	-0.020	0.858	-0.017	0.323	no
VideoKR	0.168	-0.048	0.132	0.099	0.924	-0.028	0.116	no
Video-MMMU	0.201	0.102	-0.017	0.200	0.948	-0.087	0.120	no
VideoMathQA	0.063	0.053	0.112	-0.337	0.929	-0.007	0.110	no
VideoOdyssey	0.043	–	–	–	0.824	0.007	0.286	yes
Vinoground	0.237	0.070	-0.000	1.811	0.800	0.286	0.000	no
v-HUB	0.392	0.049	-0.000	0.094	1.000	0.074	0.000	no
ALLVB	0.027	–	–	–	0.941	-0.007	0.217	yes
Causal2Needles	0.090	0.111	1.482	-0.111	0.822	0.097	0.427	no
CAViAR	0.453	0.044	-0.051	0.029	0.995	0.007	0.020	no
CG-Bench	0.233	0.214	–	0.071	–	–	0.225	no
CineTechBench	0.212	0.042	–	0.333	–	–	–	no
Colon-Bench	0.243	0.164	-0.062	0.062	0.969	-0.011	0.095	no
ContPhy	0.123	-0.150	–	0.150	–	–	–	no
E.T. Bench	0.307	0.152	-0.301	0.670	0.791	0.079	0.151	no
EgoPlan-Bench	0.004	–	–	–	–	–	0.015	yes
FineBench	0.019	–	–	–	0.964	-0.013	0.181	yes
ForeSeaQA	-0.025	–	–	–	0.967	-0.025	0.126	yes
GUI-World	0.157	0.170	-0.122	0.182	0.967	-0.086	0.057	no
HERBench	0.107	0.000	-0.219	0.344	0.925	-0.033	0.150	no
HLV-1K	0.044	–	–	–	0.905	0.115	0.270	yes
iSafetyBench	0.234	0.014	–	-0.014	–	–	–	no
K9-Bench	0.146	-0.140	0.364	-0.024	0.959	0.018	0.152	no
KFS-Bench	0.193	0.155	-0.000	0.166	0.892	0.018	0.232	no
LongVideo-Reason	0.051	0.000	0.533	-0.400	0.977	-0.014	0.125	no
M3-Verse	0.053	0.812	0.236	0.158	0.894	-0.038	0.210	no
MME-VideoOCR	0.400	0.242	0.161	0.107	0.955	0.065	0.150	no
NaturalBench-Video	0.200	0.033	-0.000	0.700	1.000	0.167	0.000	no
OVO-S-Bench	0.069	0.053	0.316	0.263	0.845	0.036	0.218	no
PLM-VideoBench	0.158	0.250	–	-0.175	–	–	0.028	no
ReactHuman	0.158	0.085	-0.000	-0.340	0.929	0.333	0.278	no
ReXTime	0.406	0.000	-0.062	0.031	0.995	-0.029	0.054	no
S-EMBER	0.107	0.531	0.719	-0.062	0.810	0.000	0.221	no
ScaleLong	0.164	0.028	0.307	0.061	0.890	0.098	0.261	no
SLVMBench	0.253	0.303	0.198	0.158	0.942	0.104	0.171	no
SmartHome-Bench	0.330	-0.010	0.054	0.013	1.000	0.049	0.022	no
TaskMeAnything-v1	0.243	0.082	-0.350	0.175	0.924	-0.053	0.117	no
TOC-Bench	0.036	–	–	–	0.927	-0.049	0.045	yes
UCF101-AD	-0.367	–	–	–	–	–	–	yes
UrbanVideo-Bench	0.090	0.259	0.039	0.039	0.914	-0.073	0.063	no
VANE-Bench	0.040	–	–	–	–	–	–	yes
VAU-Bench	0.024	–	–	–	0.996	0.013	0.025	yes
VELOCITI	0.243	0.055	–	0.164	–	–	–	no
VF-Eval	0.123	0.000	–	0.027	–	–	–	no
VGenST-Bench	0.310	0.043	–	0.204	–	–	–	no
Video-LevelGauge	0.587	0.102	-0.006	0.018	1.000	-0.014	0.038	no
VideoAds	0.377	0.009	0.106	-0.018	0.979	-0.077	0.147	no
VideoEspresso	0.214	-0.034	0.055	-0.192	1.000	0.006	0.106	no
VNBench	0.368	-0.018	0.754	0.286	0.759	0.158	0.330	no
VRU-Accident	0.177	0.094	0.377	-0.126	0.969	0.111	0.044	no

M ORDER AND TEMPORAL COVERAGE

Experiments. Four probes perturb the video along time. The shuffle probe reads the 32 uniform frames in a random order. The window probe reads 32 frames from one tenth of the video. Two half probes of the diagnostic model read 16 frames from the first or the second half. A frame-budget ladder reads 8, 32, and 128 frames, extended to 512 and 600 frames by Claude Opus 5 (Anthropic, 2026b) on the long-video set. The shuffle and window runs of Gemini 3.5 Flash (Google DeepMind, 2026b) and Qwen3-VL-8B (Bai et al., 2025) form one attacker set, and $\varepsilon_{\mathrm{order}}$ is the largest margin among them. Table 19 gives the probes per benchmark.

Settings. Gemini's shuffle uses `random.Random(42)` on 32 frames. Qwen3-VL-8B uses the per-item seed `42|qid`. The Gemini window starts uniformly in $[0, 0.9]$ and covers a tenth of the clip. Qwen3-VL-8B uses the first tenth. Both sample 32 frames within the window, or all available frames when fewer exist. Half-video probes sample 16 frames, and offset probes shift the 32-cell grid by a random phase. Frame-budget probes sample uniformly over the complete clip and use all frames when the requested count exceeds availability. Retention divides perturbed accuracy by the same model's 32-frame accuracy on matched items. Tables retain ratios above one. Figures cap values at 1.15 and identify them with hollow markers.

Results. In Table 19, the window probe reads 32 frames inside one random contiguous tenth of the video, and retention divides probe accuracy by 32-frame accuracy on the same items. Table 20 reports the Claude Opus 5 frame-budget sweep on the long-video set, the benchmarks whose sampled videos average at least 300 seconds, with the same items, prompt, and scorer and uniform frames over the whole video. Median frames is the median number of frames sent at the largest budget that ran, since a video holds at most 1,024 stored frames and short videos send every frame once. Saturation is the first budget within 5% of the maximum over the budgets that ran. The request limit ends the reference sweep at 600 frames.

Table 19: Frame-budget accuracy (%) and temporal-retention probes on 88 benchmarks

Benchmark	8f	32f	128f	Window ret.	Shuffle ret.
VNBench	41	74	90	0.37	0.71
v-HUB	70	64	69	1.70	2.15
SoMi-ToM	76	93	95	0.71	1.01
VGenST-Bench	87	88	87	0.76	0.84
MME-VideoOCR	79	84	83	0.67	0.94
UCF101-AD	45	47	46	0.94	0.97
Video-LevelGauge	85	89	89	0.80	0.98
Vinoground	80	82	78	0.72	0.66
ViMUL-Bench	82	80	76	0.79	0.90
Video-TT	51	61	64	0.55	0.97
CAViAR	77	78	78	0.96	0.99
VideoKR	47	55	66	0.66	1.06
VideoAds	74	81	85	0.78	0.95
NaturalBench-Video	83	81	82	0.74	0.66
MVBench	55	69	64	0.56	0.75
TaskMeAnything-v1	61	62	63	0.80	0.96
GasVideo-1000	64	75	84	0.76	1.09
AoTBench	68	77	69	0.51	0.70
VELOCITI	94	94	94	0.81	0.89
VRU-Accident	73	76	78	0.94	1.00
iSafetyBench	43	42	41	0.89	0.97
PAI-Bench-U	57	61	59	–	–
E.T. Bench	54	55	47	0.63	0.67
TVBench	51	63	58	0.55	0.48
VKnowU	69	71	70	0.66	0.97
VSI-Bench	43	61	–	0.39	0.54
ContPhy	58	56	58	0.87	0.88
VURB	48	62	66	0.91	1.14

Table 19 (continued)

Benchmark	8f	32f	128f	Window ret.	Shuffle ret.
SITE-Bench	48	65	62	0.60	1.00
Video-MMMU	60	70	74	0.42	0.97
TimeBlind	77	71	67	0.51	0.61
TOMATO	38	43	39	0.69	0.66
VCR-Bench	55	66	69	–	–
TempCompass	29	42	34	0.80	0.83
CineTechBench	66	60	63	0.79	0.71
Colon-Bench	72	74	75	0.85	0.96
ReactHuman	52	50	51	0.94	1.05
Video-MME	60	70	81	0.75	1.05
Video-MME-v2	26	36	46	0.63	1.28
LVBench	30	46	58	0.72	0.91
SLVMBench	68	80	81	0.81	0.97
TimeScope	26	53	63	0.35	0.90
MLVU	42	65	75	0.53	0.98
KFS-Bench	60	67	70	0.77	0.90
MMWorld	74	74	88	0.93	1.14
V2P-Bench	52	53	65	0.88	1.04
StreamingBench	47	62	69	0.76	1.00
ScaleLong	54	64	68	0.85	0.96
SmartHome-Bench	78	78	77	0.88	1.01
TOC-Bench	49	50	49	0.77	0.87
SFI-Bench	41	69	69	0.60	0.86
M3-Verse	35	40	42	0.82	0.93
PLM-VideoBench	80	85	80	0.96	0.99
UrbanVideo-Bench	51	54	52	0.81	0.88
SpatialTree	61	62	73	0.93	1.36
GUI-World	90	92	91	0.87	0.97
VANE-Bench	87	88	89	0.99	0.98
LongVideoBench	50	60	67	0.61	0.83
SAW-Bench	33	38	28	0.88	0.73
K9-Bench	70	75	78	0.83	0.97
CG-Bench	55	57	57	0.87	1.06
VideoSIAH-Eval	7	13	21	0.44	0.91
FAVOR-Bench	48	55	62	0.46	0.85
MMR-V	42	56	52	1.00	0.92
COVER	58	59	61	0.91	0.98
S-EMBER	20	37	46	0.77	0.91
HERBench	53	56	64	0.95	1.06
ReXTime	92	92	88	0.95	1.00
PerceptionComp	10	13	15	0.31	1.05
OVO-S-Bench	49	47	55	0.93	1.05
Causal2Needles	56	64	70	0.92	1.04
IntPhys 2	56	57	48	0.81	1.02
ExAct	67	69	64	0.98	0.93
TemporalBench	74	77	74	0.93	0.98
VideoEspresso	79	80	80	0.97	0.99
VSI-Super-Wild	22	32	32	0.63	0.63
ForeSeaQA	69	73	74	0.94	1.01
HLV-1K	68	69	76	0.95	0.97
DSI-Bench	70	53	74	1.21	1.27
LongVideo-Reason	74	77	79	0.96	1.03
VF-Eval	85	84	85	0.98	1.02
DSR-Bench	26	24	41	1.61	1.37
ALLVB	69	73	72	1.00	1.01
FineBench	62	62	68	0.97	0.97
HourVideo	16	21	21	0.82	0.81
VAU-Bench	96	94	95	1.00	1.01
MMSI-Video-Bench	34	28	25	0.86	1.12

Table 19 (continued)

Benchmark	8f	32f	128f	Window ret.	Shuffle ret.
EgoPlan-Bench	26	29	26	0.89	0.96
VideoMathQA	34	52	67	0.39	1.00
VideoOdyssey	28	35	42	0.69	0.94
LongTimeScope	12	16	19	0.96	1.06

Table 20: Long-video accuracy (%) at 32, 128, 512, and 600 frames with Claude Opus 5.

Benchmark	n	Chance	32f	128f	512f	600f	Med. frames	Sat.
HourVideo	300	20	25	26	26	23	600	32
LEMONADE	300	25	38	38	39	36	600	32
LongTimeScope	300	17	28	29	40	45	600	512
LongVideoBench	300	21	75	74	79	80	600	32
LVBench	300	25	57	64	74	75	600	512
MLVU	300	25	75	86	87	87	600	128
MMR-V	300	10	62	64	65	67	394	128
SciVideoBench	300	10	65	69	65	67	600	32
StreamingBench	300	25	69	75	75	76	600	128
TimeScope	300	17	64	78	89	91	600	512
V2P-Bench	300	25	57	64	70	70	360	512
Video-MME	300	25	81	83	85	85	600	32
Video-MME-v2	300	13	46	58	61	60	600	128
Video-MMMU	299	12	81	80	81	83	600	32
VideoEval-Pro	300	25	55	67	76	76	600	512
VideoKR	185	14	65	68	71	70	600	128
VideoOdyssey	320	25	49	54	61	58	600	512
VSI-Super-Wild	252	25	36	39	38	39	600	32
VURB	209	27	65	69	74	78	209	512
ALLVB	300	20	73	76	73	76	600	32
Causal2Needles	300	25	64	74	80	81	600	512
FineBench	300	25	63	63	56	56	600	32
ForeSeaQA	199	25	70	76	76	74	463	128
HERBench	300	20	58	63	63	65	360	128
HLV-1K	300	25	68	76	79	76	600	128
KFS-Bench	300	21	65	71	76	78	600	512
LongVideo-Reason	300	27	78	78	80	80	600	32
OVO-S-Bench	275	30	49	54	56	54	600	128
S-EMBER	299	20	39	48	54	49	600	512
ScaleLong	300	25	63	69	77	75	600	512

N FINE CAPABILITIES BREAK DIFFERENTLY

We resolve the first breaking level by the dominant fine capability of each benchmark within the four groups. As shown in Figure N1, the level at which a claim breaks follows the evidence the claim needs rather than the group it belongs to. Object identity breaks at the frame level on five of seven benchmarks, because a single frame already shows what the question asks about. Multi-hop reasoning breaks on text or on a frame for 15 of 29 benchmarks and survives on 6. Most reasoning questions are therefore answered by language priors or by one frame. Temporal order survives on 8 of 18 benchmarks, and four others, namely ForeSeaQA (Park et al., 2026a), LongTimeScope (Zohar et al., 2025a), TempCompass (Liu et al., 2024c), and VSI-Super-Wild (Gu et al., 2026), break on the options alone. Spatial relation breaks at the order level on 5 of 24 benchmarks and survives on 11, the highest share of any large class. A shuffled video still supports those answers, a perturbation that a claim about static relations need not resist. Fine-grained action spreads over every level and survives on 3 of 10.

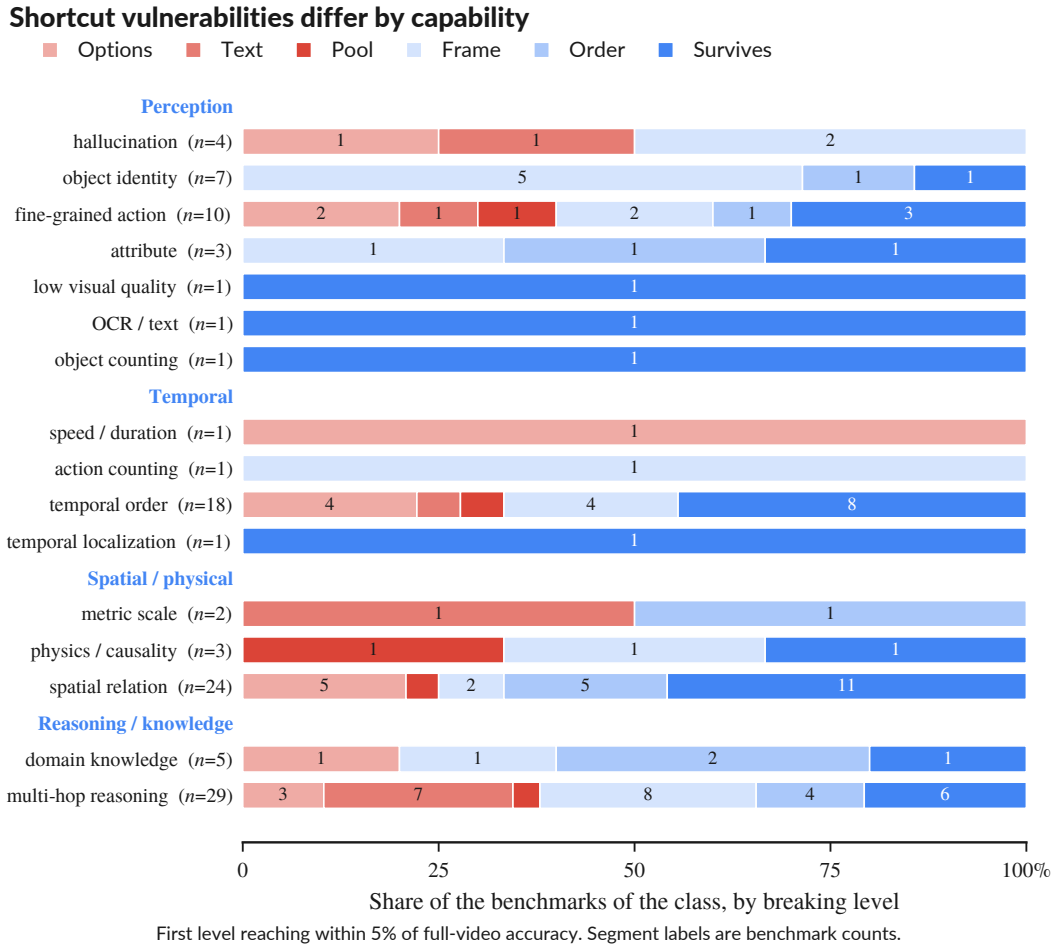


Figure N1: First breaking levels by dominant fine capability, with survivors in dark blue.

O CLAIM-MATCHED ATTACKS KEEP THE TRENDS

We recount the breaking levels with only the attacks that bear on the claim of each benchmark. Order attacks count only for the 18 benchmarks in the temporal group, because the other groups do not claim that answers depend on frame order. Blind readers do not count for the 32 reasoning and knowledge benchmarks, whose claims include knowledge that a reader brings without the video. Every other attacker, the reference, the tolerance δ, the 128-frame rescreen, and the cumulative maximum over levels stay as in the main audit, and a level left without attackers stays unmeasured. The recount reuses the item-level results of the main audit and first reproduces all 115 published breaking levels.

As shown in Table 21, the conclusions of the audit hold under claim-matched attacks. Newer benchmarks still break before visual input more often, 28% of releases from 2025 and 2026 against 15% of releases through 2024, and releases from 2025 and 2026 hold all 12 text and pool breaks. The option level is unchanged. Every order break of the main audit falls on a benchmark outside the temporal group, and these 15 benchmarks become unbroken, which raises the survivors from 38 to 53. The temporal group keeps its profile, with 6 of 18 benchmarks breaking on options and 10 surviving. Six reasoning and knowledge benchmarks move from the text level to the frame level, where frame captions alone still break them, namely COVER (Zhou et al., 2025b), LongVideo-Reason (Chen et al., 2025d), MMR-V (Zhu et al., 2026), MMWorld (He et al., 2024), VAU-Bench (Zhu et al., 2025), and VideoEspresso (Han et al., 2025). VideoSIAH-Eval (Yang et al., 2026c) moves from text to pool because its reference stays near chance, where the rule assigns the first measured level.

Table 21: Benchmarks by first breaking level under all attacks and under claim-matched attacks. All counts every attack as in the main audit. Claim drops the order level outside the temporal group and the blind readers for reasoning and knowledge. Groups hold 45 perception, 32 reasoning and knowledge, 20 spatial and physical, and 18 temporal benchmarks. Before video sums the option, text, and pool levels.

	Perception		Reasoning		Spatial		Temporal		Total	
Level	All	Claim	All	Claim	All	Claim	All	Claim	All	Claim
Option	3	3	4	4	4	4	6	6	17	17
Text	5	5	7	0	1	1	0	0	13	6
Pool	1	1	1	2	2	2	1	1	5	6
Frame	17	17	7	13	2	2	1	1	27	33
Order	6	0	5	0	4	0	0	0	15	0
Unbroken	13	19	8	13	7	11	10	10	38	53
Before video	9	9	12	6	7	7	7	7	35	29

P OTHER SETTINGS KEEP THE TRENDS

We recount the breaking levels with the tolerance δ at 3% and 8% around the 5% of the main audit, and every other setting stays fixed. As shown in Table 22, newer benchmarks break before visual input more often at every tolerance, and releases from 2025 and 2026 hold most of the text and pool breaks. A tighter tolerance leaves more benchmarks unbroken, and a looser one moves breaks to earlier levels.

Table 22: Breaking levels at three tolerances. Before video counts the option, text, and pool breaks, and the last four columns compare releases through 2024 with releases from 2025 and 2026. The 5% row is the main audit.

			Before video, share		Text and pool breaks	
δ	Before video	Unbroken	Through 2024	2025–2026	Through 2024	2025–2026
3%	33	45	19%	31%	1/26	19/89
5%	35	38	19%	34%	1/26	17/89
8%	49	19	31%	46%	4/26	21/89

We also recount the breaking levels with Claude Opus 5 (Anthropic, 2026b) at 32 frames, the strongest attacker model, as the reference for every benchmark. As shown in Table 23, this reference leaves 48 benchmarks unbroken, 31 break before visual input, and the text and pool breaks fall on the same benchmarks as in the main audit. Newer benchmarks still break before visual input more often, 29% against 19%.

Table 23: Breaking levels under the two reference models. The Gemini 3.5 Flash row is the main audit. The last two columns give the share breaking before video among releases through 2024 and from 2025 and 2026.

Reference	Option	Text	Pool	Frame	Order	Unbroken	Through 2024	2025–2026
Gemini 3.5 Flash	17	13	5	27	15	38	19%	34%
Claude Opus 5	13	12	6	22	14	48	19%	29%

We also compare the order in which benchmarks break. For every pair of benchmarks that the main audit breaks at different levels, we check whether another setting keeps their order, ties them, or reverses them. As shown in Table 24, the tolerance barely changes this order. With Claude Opus 5 as the reference, four in five pairs keep their order. With Qwen3-VL-8B (Bai et al., 2025) at 32 frames as the reference, most pairs keep their order, and the weaker reference moves most breaks to earlier levels.

Table 24: Order of breaking levels under other settings, as shares of the benchmark pairs that the main audit breaks at different levels.

Setting	Same order	Tied	Reversed
Tolerance $\delta = 3\%$	89.9%	7.8%	2.3%
Tolerance $\delta = 8\%$	80.4%	13.2%	6.5%
Claude Opus 5 reference at 32 frames	78.6%	13.3%	8.2%
Qwen3-VL-8B reference at 32 frames	57.9%	23.1%	18.9%

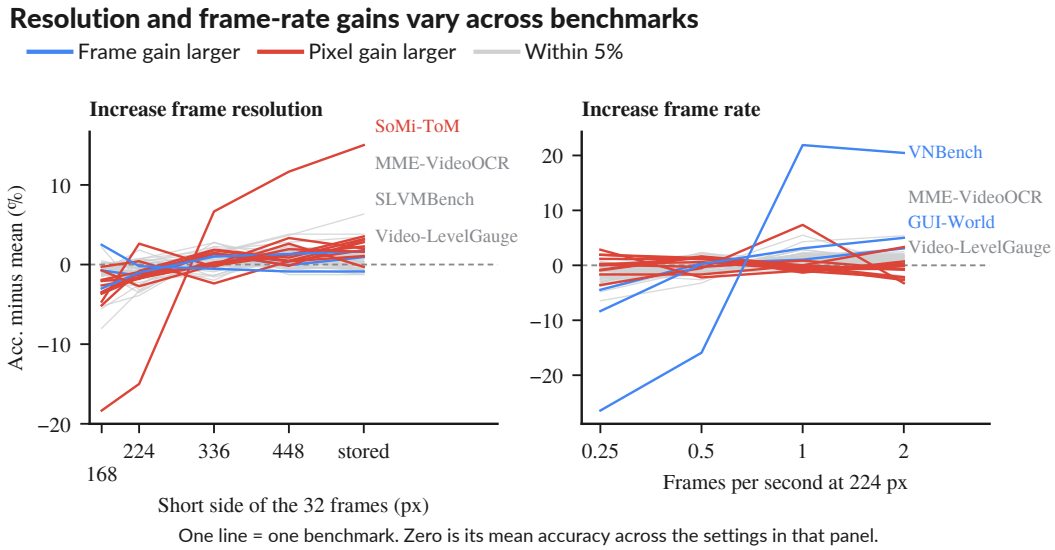


Figure Q1: Resolution and frame-rate effects, centered on each benchmark's mean accuracy.

Q FRAMES OR PIXELS

Qwen3-VL-8B (Bai et al., 2025) evaluates 300 items per benchmark passing the pool level. The resolution ladder uses 32 frames with short sides of 168, 224, 336, and 448 pixels, plus stored resolution. The temporal ladder uses 224 pixels at 0.25, 0.5, 1, and 2 frames per second, within context limits. Pixel gain is best-resolution accuracy minus accuracy at 168 pixels. Frame gain is accuracy at 2 fps minus accuracy at 0.25 fps. As shown in Figure Q1, median gains are 3.0% for pixels and 2.1% for frames. Both medians sit below the 5% tolerance. Most benchmarks respond little to either ladder, and pixels help slightly more than frames. Among the 80 benchmarks passing the pool level, 57 have both ladders, and the frame-rate ladder does not apply to 17 benchmarks of short clips.

R FRAMES CAN SUBTRACT

Video input can overturn a correct text-only answer. The blind reader answers 21% of the 9,651 attributed reference errors correctly. Across the middle half of benchmarks, this share ranges from 17 to 30%. It reaches 88% on IntPhys 2 (Bordes et al., 2025), where both conditions remain near chance but succeed on different items. Video lowers total accuracy on 10 benchmarks, including drops above 9% on LongTimeScope (Zohar et al., 2025a) and VideoOdyssey (He et al., 2026). As shown in Figure R1, video introduces individual errors on most benchmarks but lowers total accuracy on only these 10. Shortcut auditing and error attribution thus address complementary failure modes.

S NOT ALL HEADROOM IS REAL

The gap to 100% includes errors that better models cannot resolve. Of 6,863 attributed failures, 5% arise from labels. Label failures include ambiguous answers and mismatched videos. Evidence

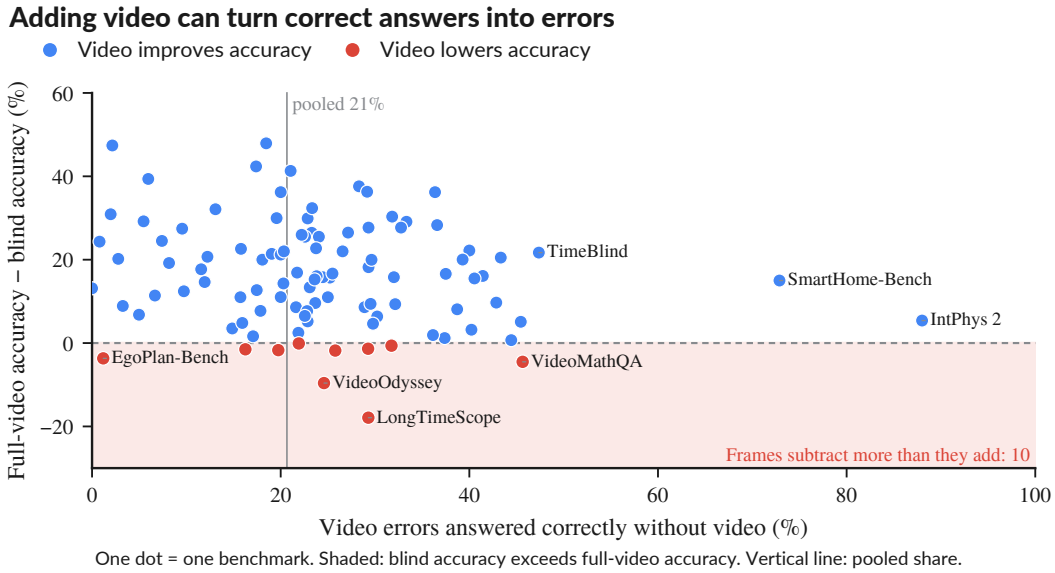


Figure R1: Blind-correct video errors versus video gain. Shading marks net losses from adding video.

missed by sparse sampling counts as a model failure. The median label-error share is 2.1%, exceeding a quarter on 1 benchmark. Its unrecoverable score gap has median 0.8% and exceeds 10% on TempCompass (Liu et al., 2024c) and UCF101-AD (Abdullah et al., 2026). Among the 38 surviving benchmarks, 30 have label-error attribution, with a median share of 0%. Of the 35 with fine attribution, language priors are the largest fine category on 30, with a median share of 54.6%. VSI-Bench (Yang et al., 2025b) instead assigns 58% to spatial relation and metric scale. As shown in Figure S1, language priors still drive most remaining errors on the survivors.

As shown in Figure S2, most named profiles concentrate their non-prior errors in two categories. Tables 25–26 give perception and temporal error shares. Tables 27–28 cover spatial, reasoning, and unresolved categories. Table 29 lists the ten benchmarks with the largest share in each category. In Tables 25 and 27, a blank cell marks a category with no instance. Benchmarks with fewer than 20 capability instances show no shares, and a dash there marks a category with at least one instance.

Table 25: Perception and temporal capability-error shares (%) across benchmarks (1/2).

Benchmark	n_{cap}	n_{det}	P-act	P-obj	P-cnt	P-ocr	P-lowq	P-hall
Main set (Gemini 3.5 Flash errors; 67 benchmarks)								
AoTBench	42	42	16.7					
Artifact-Bench	94	62	22.6	6.5		1.6	9.7	4.8
CMU	26	24	33.3					
COVER	42	39	20.5					2.6
DSI-Bench	104	96	1.0					
DSR-Bench	118	110	1.8	5.5				
EgoSchema	35	32	12.5	6.2				
ExAct	61	59	23.7					
ExpVid	38	36	38.9	8.3		16.7		
FAVOR-Bench	70	66	36.4				3.0	
GasVideo-1000	52	49	12.2				2.0	
HourVideo	129	126	1.6	5.6			0.8	
IntPhys 2	69	53		3.8				3.8
LEMONADE	91	85						
LongTimeScope	26	24				12.5		29.2
LongVideoBench	18	17	–					
LVBench	60	58		5.2	6.9			
MLVU	25	24	20.8			4.2		

Table 25 (continued)

Benchmark	n_{cap}	n_{det}	P-act	P-obj	P-cnt	P-ocr	P-lowq	P-hall
MMR-V	72	72	1.4	1.4				
MMSI-Video-Bench	128	125	3.2		2.4			
MMVU	21	20	5.0					
MMWorld	34	33	9.1	3.0				3.0
MORSE-500	35	28	3.6					
MotionBench	60	59	20.3	6.8			1.7	
MVBench	52	47	19.1	2.1				
Neptune	3	3					–	
OSI-Bench	38	38		2.6				
Perception Test	46	42	21.4	4.8				
PerceptionComp	155	140	0.7	6.4	3.6	4.3	2.1	
Q-Bench-Video	9	8					–	
ReVSI	42	42		2.4				
SAW-Bench	134	133						
SciVideoBench	35	34						
SFI-Bench	58	56		10.7	10.7	1.8		
SIS-Bench	116	111	72.1	1.8		0.9	3.6	
SITE-Bench	58	56	7.1	1.8	10.7			
SoMi-ToM	3	3				–		
SpatialTree	18	18						
ST-Evidence	20	18	5.6	5.6	5.6			
STI-Bench	132	127		0.8			0.8	
StreamingBench	41	38	2.6	5.3				
TempCompass	41	41	12.2					
TemporalBench	35	34	26.5					
TimeBlind	19	18	–					
TimeScope	20	19		5.3		10.5		10.5
TOMATO	85	84	19.0					2.4
TVBench	58	53	5.7	1.9				
v-HUB	30	26	3.8		3.8			
V2P-Bench	55	52	5.8	1.9		3.8		
Video-Holmes	47	42	2.4	2.4				
Video-MME	31	29	3.4	10.3				
Video-MME-v2	78	75	1.3	2.7	1.3		1.3	
Video-MMMU	30	25				4.0		
Video-TT	12	12	–	–				
VideoEval-Pro	66	63		7.9	1.6			
VideoKR	29	29	3.4	3.4				3.4
VideoMathQA	25	25			8.0	4.0		
VideoOdyssey	29	29		6.9				
VideoSIAH-Eval	31	29	6.9	20.7		13.8		
VidPair-Halluc	40	40		2.5				
ViMUL-Bench	15	13	–				–	
Vinoground	28	27						
VITATECS	12	12	–					
VKnowU	33	33	3.0	6.1				
VSI-Bench	36	36						2.8
VSI-Super-Wild	92	83		7.2				
VURB	26	23	8.7	8.7				
Newly onboarded benchmarks (N1) (claude-opus-5 errors; 43 benchmarks)								
ALLVB	35	35					2.9	
Causal2Needles	65	62						1.6
CAViAR	26	23					8.7	4.3
CG-Bench	23	19	10.5		5.3			
CineTechBench	22	13	7.7					
Colon-Bench	50	46	4.3	6.5				
ContPhy	43	42						
E.T. Bench	86	66	16.7	3.0				

Table 25 (continued)

Benchmark	n_{cap}	n_{det}	P-act	P-obj	P-cnt	P-ocr	P-lowq	P-hall
EgoPlan-Bench	75	66	21.2	7.6				
FineBench	30	30	13.3					
ForeSeaQA	35	32	18.8		9.4		3.1	
GUI-World	18	17	–					
HERBench	75	63	1.6	3.2	1.6			4.8
HLV-1K	46	43	2.3		4.7			
iSafetyBench	122	118	66.1				0.8	1.7
K9-Bench	44	43	34.9					
KFS-Bench	58	54	7.4	3.7		1.9		
LongVideo-Reason	33	33						
M3-Verse	121	109		2.8	0.9			1.8
MME-VideoOCR	32	32	6.2	3.1	3.1			3.1
NaturalBench-Video	19	19			–			
OVO-S-Bench	87	87		1.1	1.1			
PLM-VideoBench	29	28	7.1	3.6				
ReactHuman	120	102						
ReXTime	11	10	–					
S-EMBER	121	109	3.7	1.8	1.8	1.8	0.9	
ScaleLong	47	46	2.2	4.3	6.5			
SLVMBench	38	34				2.9		
SmartHome-Bench	30	26	23.1	7.7				
TaskMeAnything-v1	87	70	28.6	4.3				
TOC-Bench	105	102		1.0				1.0
UCF101-AD	86	83	84.3	4.8				
UrbanVideo-Bench	89	87	2.3	1.1				
VANE-Bench	26	22	9.1	4.5	4.5			4.5
VAU-Bench	7	6	–			–		
VELOCITI	13	12	–					
VF-Eval	23	18						5.6
VGenST-Bench	21	21						
Video-LevelGauge	21	19			5.3	5.3		
VideoAds	27	24			16.7			
VideoEspresso	29	28		3.6				
VNBench	38	38		2.6	10.5			
VRU-Accident	40	40						
All, primary attribution protocol (67 benchmarks)	3,413	3,205	9.1	3.0	0.9	0.9	0.8	0.6
All, Claude-role protocol (43 benchmarks)	2,170	1,993	13.1	2.0	1.4	0.3	0.3	0.7

Table 26: Perception and temporal capability-error shares (%) across benchmarks (2/2).

Benchmark	P-loc	P-attr	P-cmp	T-loc	T-cnt	T-ord	T-spd	T-cal
Main set (Gemini 3.5 Flash errors; 67 benchmarks)								
AoTBench				2.4		11.9		
Artifact-Bench								
CMU								
COVER						5.1		
DSI-Bench								
DSR-Bench				5.5			9.1	2.7
EgoSchema						3.1		
ExAct								
ExpVid								
FAVOR-Bench				3.0	1.5	6.1		
GasVideo-1000				6.1				

Table 26 (continued)

Benchmark	P-loc	P-attr	P-cmp	T-loc	T-cnt	T-ord	T-spd	T-cal
HourVideo		0.8			3.2	3.2	7.1	
IntPhys 2								
LEMONADE				17.6	4.7	12.9	25.9	
LongTimeScope								
LongVideoBench						–		
LVBench				8.6	3.4			
MLVU						4.2		
MMR-V								1.4
MMSI-Video-Bench			0.8		3.2			
MMVU								
MMWorld				3.0	6.1			
MORSE-500	3.6		3.6		7.1	7.1	3.6	
MotionBench					10.2	6.8		
MVBench				12.8	4.3	2.1		
Neptune					–			
OSI-Bench								
Perception Test					9.5			
PerceptionComp		3.6			0.7			
Q-Bench-Video								
ReVSI								
SAW-Bench			0.8					
SciVideoBench				58.8				
SFI-Bench								
SIS-Bench						0.9		
SITE-Bench	1.8			1.8	1.8	3.6		
SoMi-ToM								
SpatialTree								
ST-Evidence					11.1			
STI-Bench				1.6			3.1	0.8
StreamingBench		2.6		15.8	13.2			
TempCompass						2.4	4.9	
TemporalBench					2.9			
TimeBlind		–				–	–	
TimeScope								
TOMATO					10.7	1.2	9.5	
TVBench				15.1	22.6	5.7		
v-HUB								
V2P-Bench	1.9			1.9	5.8			
Video-Holmes						28.6		
Video-MME					6.9	6.9		
Video-MME-v2				8.0		6.7		
Video-MMMU				4.0				
Video-TT					–		–	
VideoEval-Pro					4.8			
VideoKR								
VideoMathQA								
VideoOdyssey					3.4			3.4
VideoSIAH-Eval	3.4			6.9		3.4		
VidPair-Halluc				5.0		5.0		
ViMUL-Bench					–			
Vinoground						18.5		
VITATECS						–		
VKnowU			3.0					
VSI-Bench						8.3		
VSI-Super-Wild				4.8		19.3		
VURB						4.3		
Newly onboarded benchmarks (N1) (claude-opus-5 errors; 43 benchmarks)								
ALLVB		8.6						

Table 26 (continued)

Benchmark	P-loc	P-attr	P-cmp	T-loc	T-cnt	T-ord	T-spd	T-cal
Causal2Needles		35.5						
CAViAR		8.7						
CG-Bench		10.5					5.3	
CineTechBench								
Colon-Bench		19.6						2.2
ContPhy								
E.T. Bench				15.2				10.6
EgoPlan-Bench				1.5				
FineBench								10.0
ForeSeaQA				3.1	3.1			
GUI-World						–		
HERBench	3.2	3.2	3.2		7.9	7.9		
HLV-1K	4.7			2.3		14.0		7.0
iSafetyBench								
K9-Bench		4.7				7.0		
KFS-Bench		3.7		1.9		5.6		
LongVideo-Reason						3.0		3.0
M3-Verse		2.8	7.3	5.5		1.8	5.5	
MME-VideoOCR		3.1						9.4
NaturalBench-Video								
OVO-S-Bench			2.3		1.1	4.6	4.6	
PLM-VideoBench		7.1		3.6				
ReactHuman								
ReXTime								
S-EMBER	1.8	0.9			15.6	3.7	10.1	0.9
ScaleLong		2.2			6.5			2.2
SLVMBench						8.8		
SmartHome-Bench								
TaskMeAnything-v1						1.4		
TOC-Bench	1.0		1.0	10.8	8.8	2.0	2.9	
UCF101-AD								
UrbanVideo-Bench	1.1			12.6			1.1	
VANE-Bench		4.5						
VAU-Bench								
VELOCITI								
VF-Eval		11.1						
VGenST-Bench		9.5	4.8					
Video-LevelGauge		15.8						
VideoAds		8.3						
VideoEspresso						3.6		
VNBench								
VRU-Accident		17.5						
All, primary attribution protocol (67 benchmarks)	0.1	0.2	0.1	2.9	2.3	3.0	1.8	0.2
All, Claude-role protocol (43 benchmarks)	0.4	3.5	0.7	2.2	1.8	1.9	1.3	1.0

Table 27: Spatial, reasoning, and unresolved capability-error shares (%) (1/2).

Benchmark	n_{cap}	n_{det}	S-rel	S-scale	S-phys	R-prior	R-hop
Main set (Gemini 3.5 Flash errors; 67 benchmarks)							
AoTBench	42	42				69.0	
Artifact-Bench	94	62			11.3	40.3	3.2
CMU	26	24	4.2			58.3	
COVER	42	39			2.6	53.8	7.7

Table 27 (continued)

Benchmark	n_{cap}	n_{det}	S-rel	S-scale	S-phys	R-prior	R-hop
DSI-Bench	104	96	68.8		1.0	29.2	
DSR-Bench	118	110	30.0			44.5	
EgoSchema	35	32				68.8	9.4
ExAct	61	59				69.5	
ExpVid	38	36				30.6	2.8
FAVOR-Bench	70	66				48.5	1.5
GasVideo-1000	52	49	4.1			73.5	2.0
HourVideo	129	126	13.5			62.7	0.8
IntPhys 2	69	53	20.8		49.1	22.6	
LEMONADE	91	85	3.5	10.6			
LongTimeScope	26	24				45.8	
LongVideoBench	18	17	–			–	–
LVBench	60	58	5.2			70.7	
MLVU	25	24				54.2	12.5
MMR-V	72	72				75.0	16.7
MMSI-Video-Bench	128	125	48.0	0.8		39.2	1.6
MMVU	21	20	10.0		10.0	40.0	5.0
MMWorld	34	33				69.7	
MORSE-500	35	28	17.9	3.6	25.0	21.4	3.6
MotionBench	60	59	5.1			49.2	
MVBench	52	47	10.6		6.4	42.6	
Neptune	3	3					–
OSI-Bench	38	38	47.4			50.0	
Perception Test	46	42	4.8			59.5	
PerceptionComp	155	140	43.6			16.4	17.9
Q-Bench-Video	9	8				–	
ReVSI	42	42	57.1			40.5	
SAW-Bench	134	133	60.2			39.1	
SciVideoBench	35	34				23.5	
SFI-Bench	58	56	17.9			55.4	
SIS-Bench	116	111	4.5			16.2	
SITE-Bench	58	56	25.0	7.1		35.7	3.6
SoMi-ToM	3	3				–	–
SpatialTree	18	18	–			–	
ST-Evidence	20	18				72.2	
STI-Bench	132	127	18.1	15.0		57.5	
StreamingBench	41	38	2.6			55.3	2.6
TempCompass	41	41	4.9			73.2	
TemporalBench	35	34	8.8			61.8	
TimeBlind	19	18	–			–	
TimeScope	20	19				68.4	
TOMATO	85	84	14.3			42.9	
TVBench	58	53	1.9			47.2	
v-HUB	30	26				53.8	34.6
V2P-Bench	55	52	3.8		3.8	67.3	
Video-Holmes	47	42				9.5	47.6
Video-MME	31	29				72.4	
Video-MME-v2	78	75	2.7		1.3	66.7	8.0
Video-MMMU	30	25	4.0			64.0	8.0
Video-TT	12	12				–	
VideoEval-Pro	66	63	1.6			81.0	
VideoKR	29	29			3.4	72.4	3.4
VideoMathQA	25	25	12.0			48.0	20.0
VideoOdyssey	29	29	6.9			75.9	
VideoSIAH-Eval	31	29				37.9	3.4
VidPair-Halluc	40	40	2.5			85.0	
ViMUL-Bench	15	13			–	–	
Vinoground	28	27	3.7			77.8	
VITATECS	12	12				–	

Table 27 (continued)

Benchmark	n_{cap}	n_{det}	S-rel	S-scale	S-phys	R-prior	R-hop
VKnowU	33	33	18.2			66.7	3.0
VSI-Bench	36	36	58.3			30.6	
VSI-Super-Wild	92	83				62.7	
VURB	26	23	21.7			47.8	4.3
Newly onboarded benchmarks (N1) (claude-opus-5 errors; 43 benchmarks)							
ALLVB	35	35				88.6	
Causal2Needles	65	62				62.9	
CAViAR	26	23				78.3	
CG-Bench	23	19	10.5			57.9	
CineTechBench	22	13	15.4			76.9	
Colon-Bench	50	46	2.2	4.3		56.5	
ContPhy	43	42			54.8	45.2	
E.T. Bench	86	66				54.5	
EgoPlan-Bench	75	66				30.3	36.4
FineBench	30	30				76.7	
ForeSeaQA	35	32	3.1			59.4	
GUI-World	18	17				–	–
HERBench	75	63				63.5	
HLV-1K	46	43	4.7			60.5	
iSafetyBench	122	118			0.8	30.5	
K9-Bench	44	43				53.5	
KFS-Bench	58	54				75.9	
LongVideo-Reason	33	33				84.8	9.1
M3-Verse	121	109	2.8	13.8		54.1	0.9
MME-VideoOCR	32	32	3.1			68.8	
NaturalBench-Video	19	19				–	
OVO-S-Bench	87	87	9.2	2.3		71.3	2.3
PLM-VideoBench	29	28	14.3			64.3	
ReactHuman	120	102			31.4	68.6	
ReXTime	11	10			–	–	–
S-EMBER	121	109	6.4			50.5	
ScaleLong	47	46				73.9	
SLVMBench	38	34				85.3	
SmartHome-Bench	30	26				61.5	
TaskMeAnything-v1	87	70	10.0			55.7	
TOC-Bench	105	102	1.0			69.6	1.0
UCF101-AD	86	83				9.6	
UrbanVideo-Bench	89	87	21.8			52.9	4.6
VANE-Bench	26	22				72.7	
VAU-Bench	7	6				–	
VELOCITI	13	12				–	
VF-Eval	23	18			5.6	77.8	
VGenST-Bench	21	21	19.0			66.7	
Video-LevelGauge	21	19	5.3			68.4	
VideoAds	27	24	4.2			58.3	12.5
VideoEspresso	29	28				89.3	3.6
VNBench	38	38				86.8	
VRU-Accident	40	40			2.5	77.5	
All, primary attribution protocol (67 benchmarks)	3,413	3,205	16.2	1.1	1.6	48.9	3.4
All, Claude-role protocol (43 benchmarks)	2,170	1,993	3.2	1.0	3.0	59.8	2.1

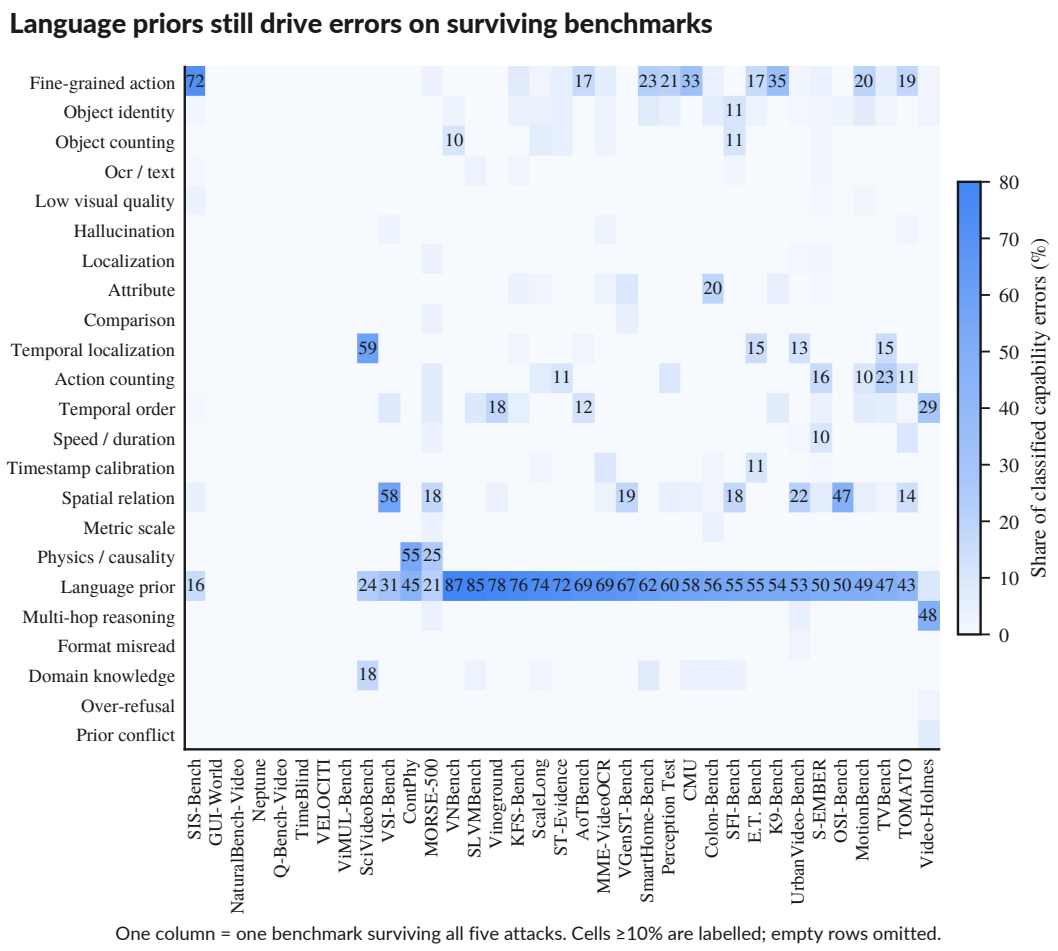


Figure S1: Fine error categories among pyramid survivors. Color encodes capability-error shares.

Table 28: Spatial, reasoning, and unresolved capability-error shares (%) (2/2).

Benchmark	R-fmt	R-know	R-refuse	R-conf	U-P	U-T	U-S	U-R
Main set (Gemini 3.5 Flash errors; 67 benchmarks)								
AoTBench								
Artifact-Bench					*30.9*		*2.1*	*1.1*
CMU		4.2			*7.7*			
COVER				7.7	*7.1*			
DSI-Bench					*6.7*		*1.0*	
DSR-Bench	0.9				*2.5*	*3.4*	*0.8*	
EgoSchema					*2.9*			*5.7*
ExAct		6.8			*3.3*			
ExpVid		2.8			*2.6*			*2.6*
FAVOR-Bench					*4.3*	*1.4*		
GasVideo-1000					*5.8*			
HourVideo	0.8				*0.8*	*1.6*		
IntPhys 2					*4.3*		*18.8*	
LEMONADE	3.5		20.0	1.2		*4.4*		*2.2*
LongTimeScope			12.5		*7.7*			
LongVideoBench					–			
LVBench						*1.7*		*1.7*
MLVU		4.2			*4.0*			
MMR-V				4.2				
MMSI-Video-Bench	0.8				*0.8*		*1.6*	

Table 28 (continued)

Benchmark	R-fmt	R-know	R-refuse	R-conf	U-P	U-T	U-S	U-R
MMVU		25.0	5.0				*4.8*	
MMWorld		3.0	3.0					*2.9*
MORSE-500					*5.7*		*8.6*	*5.7*
MotionBench					*1.7*			
MVBench					*7.7*	*1.9*		
Neptune								
OSI-Bench								
Perception Test					*8.7*			
PerceptionComp		0.7			*8.4*			*1.3*
Q-Bench-Video					–			
ReVSI								
SAW-Bench					*0.7*			
SciVideoBench		17.6						*2.9*
SFI-Bench		3.6			*3.4*			
SIS-Bench					*4.3*			
SITE-Bench					*3.4*			
SoMi-ToM								
SpatialTree	–							
ST-Evidence					*5.0*		*5.0*	
STI-Bench			2.4			*0.8*	*0.8*	*2.3*
StreamingBench					*4.9*	*2.4*		
TempCompass	2.4							
TemporalBench					*2.9*			
TimeBlind					–			
TimeScope			5.3		*5.0*			
TOMATO					*1.2*			
TVBench					*5.2*	*1.7*		*1.7*
v-HUB		3.8						*13.3*
V2P-Bench	1.9		1.9		*5.5*			
Video-Holmes			2.4	7.1				*10.6*
Video-MME					*6.5*			
Video-MME-v2					*1.3*	*1.3*		*1.3*
Video-MMMU		16.0			*6.7*			*10.0*
Video-TT								
VideoEval-Pro			1.6	1.6	*3.0*	*1.5*		
VideoKR		6.9	3.4					
VideoMathQA			8.0					
VideoOdyssey			3.4					
VideoSIAH-Eval		3.4			*6.5*			
VidPair-Halluc								
ViMUL-Bench				–	–			
Vinoground					*3.6*			
VITATECS								
VKnowU								
VSI-Bench								
VSI-Super-Wild	6.0				*1.1*	*8.7*		
VURB			4.3			*7.7*		*3.8*
Newly onboarded benchmarks (N1) (claude-opus-5 errors; 43 benchmarks)								
ALLVB								
Causal2Needles					*4.6*			
CAViAR					*11.5*			
CG-Bench					*13.0*	*4.3*		
CineTechBench					*40.9*			
Colon-Bench		4.3			*8.0*			
ContPhy							*2.3*	
E.T. Bench					*1.2*	*22.1*		
EgoPlan-Bench		3.0			*5.3*			*6.7*
FineBench								

Table 28 (continued)

Benchmark	R-fmt	R-know	R-refuse	R-conf	U-P	U-T	U-S	U-R
ForeSeaQA					*2.9*	*5.7*		
GUI-World					–			
HERBench					*16.0*			
HLV-1K					*2.2*	*4.3*		
iSafetyBench					*3.3*			
K9-Bench					*2.3*			
KFS-Bench					*6.9*			
LongVideo-Reason								
M3-Verse					*5.0*	*4.1*	*0.8*	
MME-VideoOCR								
NaturalBench-Video								
OVO-S-Bench								
PLM-VideoBench					*3.4*			
ReactHuman							*15.0*	
ReXTime						–		
S-EMBER					*5.0*	*5.0*		
ScaleLong		2.2			*2.1*			
SLVMBench		2.9			*5.3*	*5.3*		
SmartHome-Bench		7.7			*6.7*			*6.7*
TaskMeAnything-v1					*19.5*			
TOC-Bench					*1.9*	*1.0*		
UCF101-AD		1.2			*3.5*			
UrbanVideo-Bench	2.3				*1.1*	*1.1*		
VANE-Bench					*15.4*			
VAU-Bench					–			
VELOCITI					–			
VF-Eval					*21.7*			
VGenST-Bench								
Video-LevelGauge					*4.8*	*4.8*		
VideoAds					*7.4*	*3.7*		
VideoEspresso					*3.4*			
VNBench								
VRU-Accident		2.5						
All, primary attribution protocol (67 benchmarks)	0.4	0.9	1.1	0.4	*3.6*	*0.8*	*0.7*	*0.9*
All, Claude-role protocol (43 benchmarks)	0.1	0.5			*5.0*	*1.9*	*0.9*	*0.3*

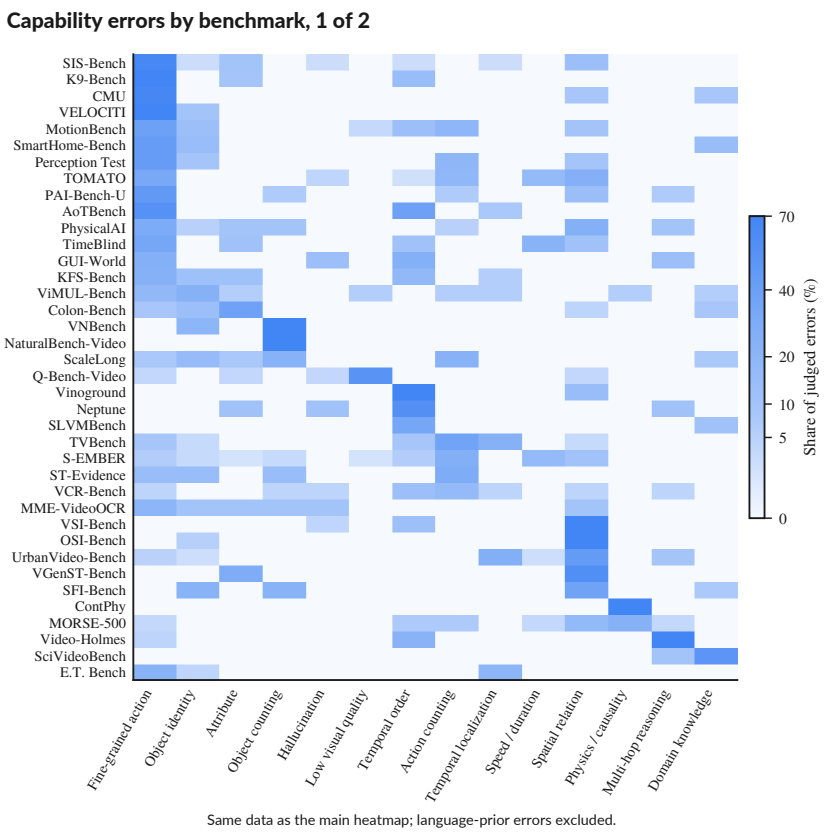
Table 29: Top ten benchmarks per fine error category (% share); at least 20 capability errors required.

Category	Benchmarks
Perception	
fine-grained action	SIS-Bench 69, ExpVid 37, FAVOR-Bench 34, CMU 31, TemporalBench 26, ExAct 23, MLVU 20, MotionBench 20, Perception Test 20, COVER 19
object identity	VideoSIAH-Eval 19, SFI-Bench 10, Video-MME 10, ExpVid 8, VURB 8, VideoEval-Pro 8, VideoOdyssey 7, MotionBench 7, VSI-Super-Wild 7, VKnowU 6
object counting	SFI-Bench 10, SITE-Bench 10, VideoMathQA 8, LVBench 7, ST-Evidence 5, v-HUB 3, PerceptionComp 3, MMSI-Video-Bench 2, VideoEval-Pro 2, Video-MME-v2 1
text and OCR	ExpVid 16, VideoSIAH-Eval 13, LongTimeScope 12, TimeScope 10, MLVU 4, VideoMathQA 4, PerceptionComp 4, V2P-Bench 4, Video-MMMU 3, SFI-Bench 2

Table 29 (continued)

Category	Benchmarks
low visual quality	Artifact-Bench 6, SIS-Bench 3, FAVOR-Bench 3, PerceptionComp 2, GasVideo-1000 2, MotionBench 2, Video-MME-v2 1, HourVideo 1, STI-Bench 1
hallucinated content	LongTimeScope 27, TimeScope 10, VideoKR 3, Artifact-Bench 3, MMWorld 3, IntPhys 2 3, VSI-Bench 3, COVER 2, TOMATO 2
localization	VideoSIAH-Eval 3, MORSE-500 3, V2P-Bench 2, SITE-Bench 2
attribute	PerceptionComp 3, StreamingBench 2, HourVideo 1
comparison	VKnowU 3, MORSE-500 3, MMSI-Video-Bench 1, SAW-Bench 1
Temporal	
temporal localization	SciVideoBench 57, LEMONADE 16, StreamingBench 15, TVBench 14, MVBench 12, LVBench 8, Video-MME-v2 8, VideoSIAH-Eval 6, GasVideo-1000 6, DSR-Bench 5
action counting	TVBench 21, StreamingBench 12, TOMATO 11, ST-Evidence 10, MotionBench 10, Perception Test 9, Video-MME 6, MMWorld 6, MORSE-500 6, V2P-Bench 5
temporal order	Video-Holmes 26, Vinoground 18, VSI-Super-Wild 17, LEMONADE 12, AoTBench 12, VSI-Bench 8, MotionBench 7, Video-MME 6, Video-MME-v2 6, FAVOR-Bench 6
speed and duration	LEMONADE 24, TOMATO 9, DSR-Bench 8, HourVideo 7, TempCompass 5, STI-Bench 3, MORSE-500 3
frame-to-timestamp calibration	VideoOdyssey 3, DSR-Bench 3, MMR-V 1, STI-Bench 1
Spatial and physical	
spatial relation	DSI-Bench 63, SAW-Bench 60, VSI-Bench 58, ReVSI 57, OSI-Bench 47, MMSI-Video-Bench 47, PerceptionComp 39, DSR-Bench 28, SITE-Bench 24, VURB 19
metric scale	STI-Bench 14, LEMONADE 10, SITE-Bench 7, MORSE-500 3, MMSI-Video-Bench 1
physics and causality	IntPhys 2 38, MORSE-500 20, MMVU 10, Artifact-Bench 7, MVBench 6, V2P-Bench 4, VideoKR 3, COVER 2, Video-MME-v2 1, DSI-Bench 1
Reasoning and knowledge	
language prior	VidPair-Halluc 85, VideoEval-Pro 77, VideoOdyssey 76, MMR-V 75, Vinoground 75, TempCompass 73, VideoKR 72, GasVideo-1000 69, AoTBench 69, LVBench 68
multi-hop reasoning	Video-Holmes 43, v-HUB 30, VideoMathQA 20, MMR-V 17, PerceptionComp 16, MLVU 12, EgoSchema 9, Video-MME-v2 8, COVER 7, Video-MMMU 7
task format	VSI-Super-Wild 5, LEMONADE 3, TempCompass 2, V2P-Bench 2, DSR-Bench 1, HourVideo 1, MMSI-Video-Bench 1
domain knowledge	MMVU 24, SciVideoBench 17, Video-MMMU 13, VideoKR 7, ExAct 7, MLVU 4, CMU 4, SFI-Bench 3, v-HUB 3, VideoSIAH-Eval 3
over-refusal	LEMONADE 19, LongTimeScope 12, VideoMathQA 8, TimeScope 5, MMVU 5, VURB 4, VideoKR 3, VideoOdyssey 3, MMWorld 3, STI-Bench 2
language prior conflict	COVER 7, Video-Holmes 6, MMR-V 4, VideoEval-Pro 2, LEMONADE 1

Tables 30–31 report attribution shares and exclusion rates from the error-verification workflow. Tables 32–33 give the Claude-role analysis on newly onboarded benchmarks. Tables 30 and 32 count errors as the attributed 32-frame failures of Gemini 3.5 Flash (Google DeepMind, 2026b) and of



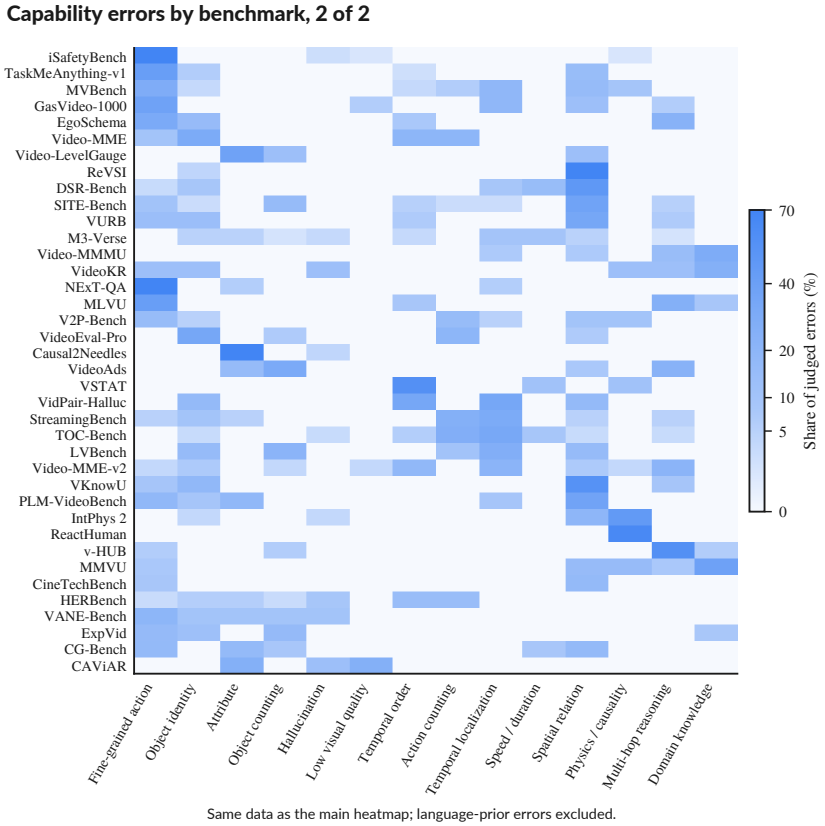


Figure S2: Named capability-error profiles, first half above and second half below. Percentages of judged non-prior errors.

Claude Opus 5 (Anthropic, 2026b) in turn, and the attributed count removes the disputed, unstable, and unverified instances to form the denominator of the three shares. Per-benchmark rows require at least 20 errors, which keeps 63 of 77 benchmarks in the first table and 40 of 43 in the second. In the second table the caption signal shown to the judges is answered by the erring model itself. Five rows of the first table, SIS-Bench (Zou et al., 2026), Video-Holmes (Cheng et al., 2025), SciVideoBench (Deng et al., 2025), STI-Bench (Li et al., 2025), and ExpVid (Xu et al., 2025), were computed from predictions made with letter-only option lists.

Table 30: Verified attribution shares and excluded-instance rates (%) (1/2).

Benchmark	Errors	Attrib.	Annot.	Cover.
Pooled, instance-weighted	6,114	4,618	21.4	4.6
Pooled, item-weighted	6,114	4,618	21.4	4.6
LongTimeScope	198	114	75.4	1.8
SciVideoBench	282	151	74.2	2.6
Video-Holmes	250	159	69.2	1.3
ExpVid	191	116	67.2	0.0
TempCompass	112	101	59.4	0.0
VideoSIAH-Eval	184	117	48.7	24.8
LEMONADE	257	183	47.5	2.7
TimeScope	120	73	45.2	27.4
SIS-Bench	279	176	34.1	0.0
Video-MMMU	59	45	33.3	0.0
VKnowU	54	50	32.0	2.0
SpatialTree	27	24	25.0	0.0
MMWorld	49	44	22.7	0.0
Q-Bench-Video	22	14	21.4	14.3
LongVideoBench	49	33	21.2	24.2
COVER	79	53	20.8	0.0
V2P-Bench	119	81	19.8	12.3
MMR-V	111	93	19.4	3.2
StreamingBench	73	57	17.5	10.5
HourVideo	203	159	17.0	1.9
Video-MME	58	42	16.7	9.5
ViMUL-Bench	26	20	15.0	10.0
LVBench	115	87	14.9	16.1
VURB	49	34	14.7	8.8
VideoOdyssey	61	48	14.6	25.0
EgoSchema	53	42	14.3	2.4
STI-Bench	207	151	11.9	0.7
VideoEval-Pro	131	93	11.8	17.2
Video-MME-v2	121	97	11.3	8.2
VidPair-Halluc	52	45	11.1	0.0
MLVU	70	46	10.9	34.8
VideoKR	45	32	9.4	0.0
VideoMathQA	46	32	9.4	12.5
MMSI-Video-Bench	159	137	6.6	0.0
SITE-Bench	70	63	6.3	1.6
v-HUB	40	32	6.2	0.0
PerceptionComp	221	178	5.6	7.3
MVBench	70	55	5.5	0.0
ST-Evidence	27	21	4.8	0.0
MMVU	27	22	4.5	0.0
FAVOR-Bench	90	73	4.1	0.0
GasVideo-1000	63	55	3.6	1.8
SFI-Bench	69	60	3.3	0.0
TOMATO	126	92	3.3	4.3
ExAct	70	63	3.2	0.0
MORSE-500	47	36	2.8	0.0
TemporalBench	44	36	2.8	0.0

Table 30 (continued)

Benchmark	Errors	Attrib.	Annot.	Cover.
Perception Test	58	47	2.1	0.0
Artifact-Bench	107	96	2.1	0.0
DSR-Bench	146	120	1.7	0.0
TVBench	79	66	1.5	10.6
VSI-Super-Wild	110	93	1.1	0.0
DSI-Bench	107	105	0.9	0.0
AoTBench	48	42	0.0	0.0
CMU	34	26	0.0	0.0
IntPhys 2	100	76	0.0	9.2
MotionBench	70	62	0.0	3.2
OSI-Bench	39	38	0.0	0.0
ReVSI	45	43	0.0	2.3
SAW-Bench	140	134	0.0	0.0
TimeBlind	23	19	0.0	0.0
Vinoground	33	28	0.0	0.0
VSI-Bench	40	36	0.0	0.0

Table 31: Verified attribution shares and excluded-instance rates (%) (2/2).

Benchmark	Capab.	Disput.	Unstab.	Unverif.
Pooled, instance-weighted	74.0	19.4	1.2	3.9
Pooled, item-weighted	74.0	19.4	1.2	3.9
LongTimeScope	22.8	29.3	1.0	12.1
SciVideoBench	23.2	41.1	0.0	5.3
Video-Holmes	29.6	33.6	0.0	2.8
ExpVid	32.8	35.6	0.0	3.7
TempCompass	40.6	9.8	0.0	0.0
VideoSIAH-Eval	26.5	15.8	0.0	20.7
LEMONADE	49.7	25.7	0.0	3.1
TimeScope	27.4	27.5	3.3	8.3
SIS-Bench	65.9	36.6	0.4	0.0
Video-MMMU	66.7	23.7	0.0	0.0
VKnowU	66.0	7.4	0.0	0.0
SpatialTree	75.0	7.4	3.7	0.0
MMWorld	77.3	6.1	4.1	0.0
Q-Bench-Video	64.3	31.8	0.0	4.5
LongVideoBench	54.5	20.4	4.1	8.2
COVER	79.2	26.6	0.0	6.3
V2P-Bench	67.9	22.7	0.8	8.4
MMR-V	77.4	14.4	0.0	1.8
StreamingBench	71.9	17.8	2.7	1.4
HourVideo	81.1	14.3	1.5	5.9
Video-MME	73.8	19.0	5.2	3.4
ViMUL-Bench	75.0	23.1	0.0	0.0
LVBench	69.0	13.9	1.7	8.7
VURB	76.5	20.4	4.1	6.1
VideoOdyssey	60.4	11.5	0.0	9.8
EgoSchema	83.3	18.9	0.0	1.9
STI-Bench	87.4	27.1	0.0	0.0
VideoEval-Pro	71.0	8.4	3.1	17.6
Video-MME-v2	80.4	12.4	0.8	6.6
VidPair-Halluc	88.9	9.6	0.0	3.8
MLVU	54.4	25.7	1.4	7.1
VideoKR	90.6	26.7	0.0	2.2
VideoMathQA	78.1	19.6	4.3	6.5

Table 31 (continued)

Benchmark	Capab.	Disput.	Unstab.	Unverif.
MMSI-Video-Bench	93.4	8.8	1.3	3.8
SITE-Bench	92.1	7.1	1.4	1.4
v-HUB	93.8	15.0	5.0	0.0
PerceptionComp	87.1	18.1	0.0	1.4
MVBench	94.5	17.1	4.3	0.0
ST-Evidence	95.2	14.8	0.0	7.4
MMVU	95.5	18.5	0.0	0.0
FAVOR-Bench	95.9	13.3	3.3	2.2
GasVideo-1000	94.5	7.9	1.6	3.2
SFI-Bench	96.7	10.1	2.9	0.0
TOMATO	92.4	21.4	4.0	1.6
ExAct	96.8	10.0	0.0	0.0
MORSE-500	97.2	23.4	0.0	0.0
TemporalBench	97.2	9.1	6.8	2.3
Perception Test	97.9	15.5	3.4	0.0
Artifact-Bench	97.9	10.3	0.0	0.0
DSR-Bench	98.3	16.4	1.4	0.0
TVBench	87.9	13.9	1.3	1.3
VSI-Super-Wild	98.9	15.5	0.0	0.0
DSI-Bench	99.1	1.9	0.0	0.0
AoTBench	100.0	6.2	6.2	0.0
CMU	100.0	23.5	0.0	0.0
IntPhys 2	90.8	16.0	2.0	6.0
MotionBench	96.8	5.7	0.0	5.7
OSI-Bench	100.0	2.6	0.0	0.0
ReVSI	97.7	0.0	4.4	0.0
SAW-Bench	100.0	3.6	0.7	0.0
TimeBlind	100.0	8.7	8.7	0.0
Vinoground	100.0	12.1	3.0	0.0
VSI-Bench	100.0	7.5	2.5	0.0

Table 32: Claude-role attribution and exclusion rates on new benchmarks (%) (1/2).

Benchmark	Errors	Attrib.	Annot.	Cover.
Pooled, instance-weighted	3,058	2,322	5.0	2.2
Pooled, item-weighted	3,058	2,322	5.0	2.2
CAViAR	60	42	38.1	0.0
SmartHome-Bench	59	42	28.6	0.0
VideoAds	44	35	22.9	0.0
UCF101-AD	151	109	21.1	0.0
VideoEspresso	47	33	12.1	0.0
EgoPlan-Bench	194	85	11.8	0.0
ScaleLong	89	58	10.3	8.6
ALLVB	63	43	9.3	9.3
VRU-Accident	58	44	9.1	0.0
LongVideo-Reason	57	36	8.3	0.0
CG-Bench	43	26	7.7	3.9
KFS-Bench	84	67	7.5	6.0
SLVMBench	51	41	7.3	0.0
GUI-World	24	19	5.3	0.0
iSafetyBench	157	127	3.9	0.0
VNBench	65	51	3.9	21.6
HLV-1K	83	53	3.8	9.4
MME-VideoOCR	38	33	3.0	0.0
OVO-S-Bench	123	93	2.1	4.3

Table 32 (continued)

Benchmark	Errors	Attrib.	Annot.	Cover.
UrbanVideo-Bench	126	90	1.1	0.0
TOC-Bench	119	106	0.9	0.0
M3-Verse	148	122	0.8	0.0
Causal2Needles	86	65	0.0	0.0
CineTechBench	27	22	0.0	0.0
Colon-Bench	56	50	0.0	0.0
ContPhy	50	43	0.0	0.0
E.T. Bench	99	86	0.0	0.0
FineBench	41	32	0.0	6.2
ForeSeaQA	44	35	0.0	0.0
HERBench	114	79	0.0	5.1
K9-Bench	56	44	0.0	0.0
NaturalBench-Video	22	21	0.0	9.5
PLM-VideoBench	31	29	0.0	0.0
ReactHuman	126	120	0.0	0.0
S-EMBER	168	129	0.0	6.2
TaskMeAnything-v1	98	87	0.0	0.0
VANE-Bench	30	26	0.0	0.0
VF-Eval	35	25	0.0	8.0
VGenST-Bench	23	21	0.0	0.0
Video-LevelGauge	24	21	0.0	0.0

Table 33: Claude-role attribution and exclusion rates on new benchmarks (%) (2/2).

Benchmark	Capab.	Disput.	Unstab.	Unverif.
Pooled, instance-weighted	92.7	19.9	1.1	3.0
Pooled, item-weighted	92.7	19.9	1.1	3.0
CAViAR	61.9	20.0	1.7	8.3
SmartHome-Bench	71.4	28.8	0.0	0.0
VideoAds	77.1	18.2	2.3	0.0
UCF101-AD	78.9	27.8	0.0	0.0
VideoEspresso	87.9	23.4	0.0	6.4
EgoPlan-Bench	88.2	34.0	0.0	22.2
ScaleLong	81.0	27.0	1.1	6.7
ALLVB	81.4	27.0	0.0	4.8
VRU-Accident	90.9	22.4	0.0	1.7
LongVideo-Reason	91.7	29.8	3.5	3.5
CG-Bench	88.5	32.6	0.0	7.0
KFS-Bench	86.6	13.1	3.6	3.6
SLVMBench	92.7	17.6	2.0	0.0
GUI-World	94.7	16.7	0.0	4.2
iSafetyBench	96.1	15.3	0.6	3.2
VNBench	74.5	16.9	1.5	3.1
HLV-1K	86.8	33.7	0.0	2.4
MME-VideoOCR	97.0	5.3	2.6	5.3
OVO-S-Bench	93.5	22.8	0.8	0.8
UrbanVideo-Bench	98.9	28.6	0.0	0.0
TOC-Bench	99.1	10.9	0.0	0.0
M3-Verse	99.2	15.5	2.0	0.0
Causal2Needles	100.0	22.1	1.2	1.2
CineTechBench	100.0	18.5	0.0	0.0
Colon-Bench	100.0	10.7	0.0	0.0
ContPhy	100.0	14.0	0.0	0.0
E.T. Bench	100.0	10.1	2.0	1.0
FineBench	93.8	19.5	2.4	0.0

Table 33 (continued)

Benchmark	Capab.	Disput.	Unstab.	Unverif.
ForeSeaQA	100.0	15.9	2.3	2.3
HERBench	94.9	27.2	3.5	0.0
K9-Bench	100.0	17.9	3.6	0.0
NaturalBench-Video	90.5	4.5	0.0	0.0
PLM-VideoBench	100.0	6.5	0.0	0.0
ReactHuman	100.0	4.0	0.0	0.8
S-EMBER	93.8	18.5	2.4	2.4
TaskMeAnything-v1	100.0	11.2	0.0	0.0
VANE-Bench	100.0	10.0	3.3	0.0
VF-Eval	92.0	25.7	0.0	2.9
VGenST-Bench	100.0	4.3	4.3	0.0
Video-LevelGauge	100.0	12.5	0.0	0.0

All 1,826 yes-or-no items in VF-Eval (Song et al., 2025) have answer yes, and 282 CAViAR (Garg et al., 2026) items carry conflicting labels. Blind accuracy falls 11 to 21% below chance on 4 sources, namely VideoSIAH-Eval (Yang et al., 2026c), VidPair-Halluc (Huang et al., 2026a), PerceptionComp (Li et al., 2026c), and Artifact-Bench (Tang et al., 2026a). These results flag formats whose question-answer relations mislead the blind reader. Chance metadata also needs checking. 33 multiple-choice benchmarks omit it, and sampled option counts differ by at least 5% on 13 of 15 with declarations. Micro versus macro averaging reverses 79 of 1,344 system pairs. Eight benchmarks are flagged by three of the four independent screens. CAViAR, ExAct (Yi et al., 2025), LongTimeScope (Zohar et al., 2025a), TimeScope (Zohar et al., 2025b), and UCF101-AD break at or before the frame level, iSafetyBench (Abdullah et al., 2025) breaks at the order level, and ColonBench (Hamdi et al., 2026) and SmartHome-Bench (Zhao et al., 2025b) survive. Those screens assess protocol errors, effective size, near-duplicates, and text predictability.

T WHICH BENCHMARKS TO USE FOR A CAPABILITY

A benchmark that survives every level holds a certificate that none of our attackers can forge. Of the 115 benchmarks with a certificate, 38 survive. For perception, 13 of 45, namely ColonBench (Hamdi et al., 2026), K9-Bench (Attarde et al., 2026), KFS-Bench (Li et al., 2026f), MME-VideoOCR (Shi et al., 2025), NaturalBench-Video (Li, 2025), Perception Test (Patraucean et al., 2023), Q-Bench-Video (Zhang et al., 2025f), ST-Evidence (Wang et al., 2026b), ScaleLong (Ma et al., 2025), TimeBlind (Li et al., 2026a), VELOCITI (Saravanan et al., 2025), VNBench (Zhao et al., 2025d), and ViMUL-Bench (Shafique et al., 2025). For temporal understanding, 10 of 18, namely AoTBench (Xue et al., 2025), E.T. Bench (Liu et al., 2024b), MotionBench (Hong et al., 2025), PhysicalAI (Zhou et al., 2026), S-EMBER (Wang et al., 2026c), SLVMBench (Yang et al., 2026b), TOMATO (Shangguan et al., 2024), TVBench (Cores et al., 2024), VCR-Bench (Qi et al., 2025), and Vinoground (Zhang et al., 2024a). For spatial and physical understanding, 7 of 20, namely CMU (Tan et al., 2026), ContPhy (Zheng et al., 2024b), OSI-Bench (Wu et al., 2026), SIS-Bench (Zou et al., 2026), UrbanVideo-Bench (Zhao et al., 2025a), VGenST-Bench (Park et al., 2026b), and VSI-Bench (Yang et al., 2025b). For reasoning and knowledge, 8 of 32, namely GUI-World (Chen et al., 2025a), MORSE-500 (Cai et al., 2025), Neptune (Nagrani et al., 2024), PAI-Bench-U (Zhou et al., 2026), SFI-Bench (Zhang et al., 2026a), SciVideoBench (Deng et al., 2025), SmartHome-Bench (Zhao et al., 2025b), and Video-Holmes (Cheng et al., 2025). A survivor resists the attackers we tried and certifies nothing beyond them. Its report card names the level at which it came closest to breaking, and a user who needs the strongest guard for one claim can rank the survivors of a group by that margin. Tables 34, 35, 36, and 37 give that ranking for the ten strongest benchmarks of each group, where the margin is the breaking threshold $s^* - c - \delta$ minus the largest exploitability over the measured levels, in %.

U POOL CONSTRUCTION

Options-only and blind screens use Qwen3-VL-8B (Bai et al., 2025) margins over chance across four option permutations. The single-frame screen uses the correct option's log-probability margin. The

Table 34: Perception. Ten strongest of 45 benchmarks, 13 of which survive, ranked by breaking level and then by the margin below the breaking threshold (%).

Rank	Benchmark	Level	Margin
1	VNBench	Survives	+17.1
2	NaturalBench-Video	Survives	+15.7
3	MME-VideoOCR	Survives	+12.3
4	Colon-Bench	Survives	+5.7
5	ViMUL-Bench	Survives	+5.5
6	Q-Bench-Video	Survives	+3.3
7	TimeBlind	Survives	+3.2
8	K9-Bench	Survives	+2.8
9	KFS-Bench	Survives	+2.3
10	ScaleLong	Survives	+1.4

Table 35: Temporal. Ten strongest of 18 benchmarks, 10 of which survive, ranked by breaking level and then by the margin below the breaking threshold (%).

Rank	Benchmark	Level	Margin
1	VCR-Bench	Survives	+9.8
2	TVBench	Survives	+8.2
3	PhysicalAI	Survives	+8.1
4	TOMATO	Survives	+6.4
5	MotionBench	Survives	+3.8
6	E.T. Bench	Survives	+1.3
7	SLVMBench	Survives	+1.3
8	AoTBench	Survives	+1.2
9	S-EMBER	Survives	+0.7
10	Vinoground	Survives	+0.5

Table 36: Spatial / physical. Ten strongest of 20 benchmarks, 7 of which survive, ranked by breaking level and then by the margin below the breaking threshold (%).

Rank	Benchmark	Level	Margin
1	VGenST-Bench	Survives	+12.0
2	SIS-Bench	Survives	+5.3
3	CMU	Survives	+4.1
4	ContPhy	Survives	+3.6
5	VSI-Bench	Survives	+2.7
6	OSI-Bench	Survives	+1.0
7	UrbanVideo-Bench	Survives	+0.0
8	ReVSI	Order	-3.3
9	M3-Verse	Order	-3.4
10	VURB	Order	-7.3

Table 37: Reasoning / knowledge. Ten strongest of 32 benchmarks, 8 of which survive, ranked by breaking level and then by the margin below the breaking threshold (%).

Rank	Benchmark	Level	Margin
1	PAI-Bench-U	Survives	+6.0
2	Video-Holmes	Survives	+4.7
3	Neptune	Survives	+4.0
4	SciVideoBench	Survives	+2.3
5	SFI-Bench	Survives	+2.1
6	SmartHome-Bench	Survives	+1.7
7	MORSE-500	Survives	+1.0
8	GUI-World	Survives	+0.7
9	EgoSchema	Order	-5.4
10	Video-MMMU	Order	-5.8

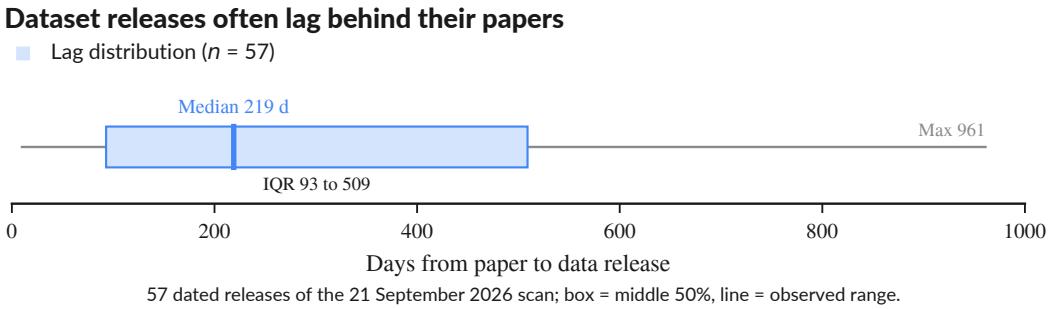


Figure U1: Paper-to-data lag of the 57 dated candidates at the scan of 21 September 2026.

32-frame and shuffled-frame screens use Qwen3-VL-2B answers, followed by cross-source question deduplication. Records retain source, video identifier, duration, task, event, format, option count, license, attack margins, and model outcomes. They also retain BGE-large (Xiao et al., 2024) question embeddings and SigLIP2 (Tschannen et al., 2025) middle-frame embeddings. An item labeler assigns 18 fine capability categories in four groups from the question, options, and gold answer. A scene labeler uses eight uniform frames and the source name, without question text, to assign one of 19 video-content categories, among them daily life, sports, films, scientific experiments, and an "other" class. We retain at least one item per available task-event cell and flag cells below 30 items as thin. Caption attacks run on selected candidates because captions depend on the requested frame sample.

The watcher scans registries, technical reports, and release pages every twelve hours, with a retrospective scan over 32 months. At the scan of 21 September 2026, the watcher lists 585 candidates, 69 of them in the pyramid. As shown in Figure U1, the median paper-to-data lag over 57 dated releases is 219 days and the middle half spans 93 to 509 days. At that scan, 44 candidates had released only part of their videos or annotations. Each approved source is pinned to a revision and audited before joining a pool snapshot. Compositions can be reissued on later snapshots. As shown in Figure U2, household and daily life supplies the most items and surveillance and security anomalies the fewest, 894, and even the rarest class supplies candidates for targeted compositions.

A composition query uses equality and range predicates over task, event, duration, format, option count, and license, with coverage quotas when requested. The lockfile records the round-robin selector's caps and any relaxations. The gate applies each attack's item-level success rule. Options-only and blind checks use margins across four option permutations. Single-frame and caption checks use answer correctness. Order checks use shuffled frames or the first tenth. Refilling stops when no item is dropped. The JSON lockfile records snapshot hash, query, selector version, seed, item identifiers, option permutations, round count, and residual exploitability. Its SHA-256 hash names the released benchmark.

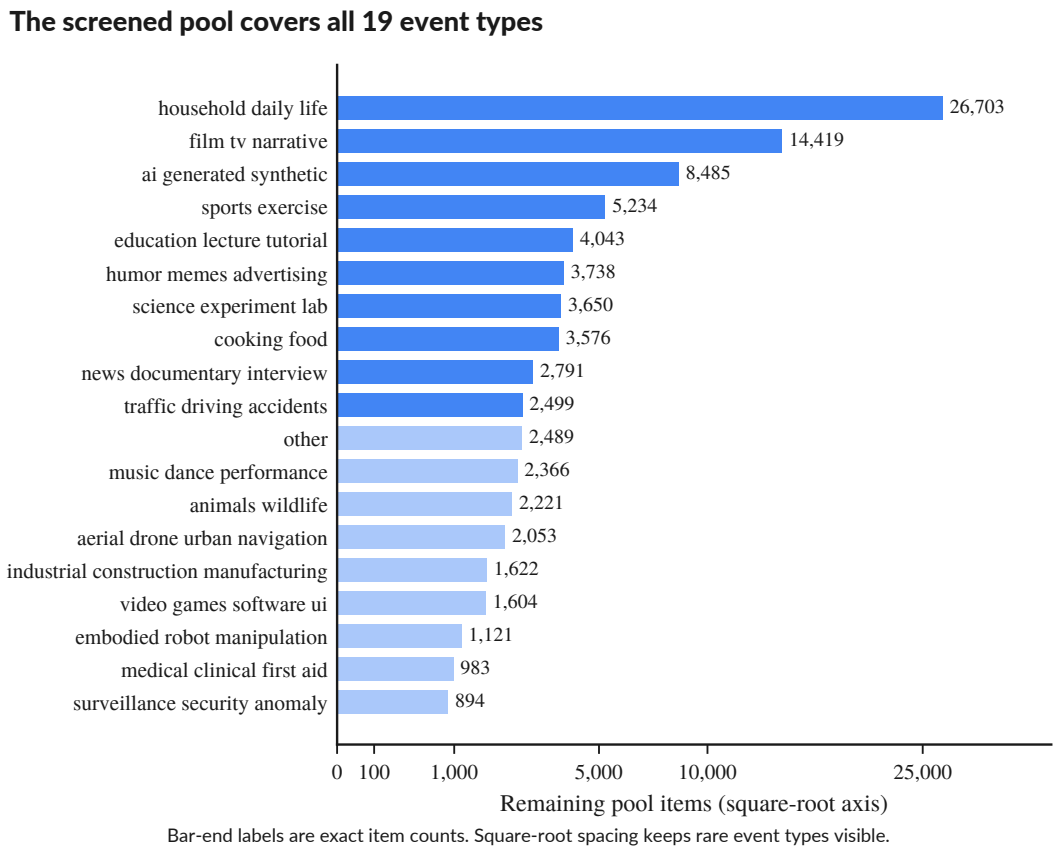


Figure U2: Screened QA counts across nineteen event types. Bar labels give exact counts.

We release Video-Index alongside other meta-benchmarks that prioritize coverage breadth, namely the 10,000-item coverage composition and a separate 834-item coverage census that draws up to eight items from each source without attacker filtering. Tables 38 to 46 give, in turn, Video-Index accuracy by duration, viewpoint, and authenticity, the census scores and video gains, the scene clusters by duration, the coverage composition by scene and duration, and each source's contribution against its attack results.

Collecting the meta-benchmark. The meta-benchmark is a difficulty-first census of the screened pool. It contains 840 items and uses the following construction procedure.

Candidates and attacker difficulty. We start from 63,630 multiple-choice items from the 115 audited benchmarks. Every screening step is complete, and no declared attacker solves these items. The attackers receive options only, question text, a single frame, 32 frames with the 2B model, or shuffled frames. For each item, we rank each of its five attacker margins by percentile within the pool. The largest percentile is the item's worst-case attacker percentile. A small value means that even the strongest attacker remains far from the answer. We retain the 3,000 lowest-ranked items, with one item per video and ties broken by item identifier.

Capability labels. Claude Opus 5 (Anthropic, 2026b) reads the question, options, and gold answer without frames. It assigns the primary capability needed to answer from video, choosing among the 18 capability categories of the attribution taxonomy. The five data-and-protocol categories are withheld. Each category belongs to Perception, Temporal, Spatial / physical, or Reasoning / knowledge. The same call flags first-person wording and questions about presence or authenticity. These flags and video duration replace the former long-video, egocentric and hallucination classes. We report them as appendix attributes.

Answer verification. Selecting items that defeat every attacker also concentrates questionable gold answers. We therefore check every candidate before selection. Claude Opus 5 receives the question, options, marked answer and video frames sampled at 1 fps, initially capped at 32. Items not confirmed at that budget are reread with the evaluation budget of 512 frames, and the later verdict takes precedence.

The verifier returns one of four verdicts. Supported means that the frames establish the marked answer and rule out the alternatives. Contradicted means that another option is visibly correct. Definition dependent means that the answer depends on a benchmark-specific definition or subjective judgment.

We keep supported items and reversed clips from AoTBench (Xue et al., 2025) RTime and TempCompass (Liu et al., 2024c) direction, whose labels describe the reversed action by construction. We drop contradicted and definition-dependent items, and not-verifiable items from shorter videos. Two format checks precede verification. VidPair-Halluc ordering questions whose choices are individual events and whose stored answer is a single letter cannot be scored as single-choice items. We exclude them. Options stored as label/text records are rendered as text.

Selection by attacker difficulty. Within each of the four capability groups, we take verified items in increasing worst-case attacker percentile, with one item per video, until reaching 210. Reference-model accuracy does not enter the selection rule.

The final census contains 840 items from 76 benchmarks. The worst-case attacker percentile has median 0.14 and maximum 0.23. Mean chance is 29.7 %. The largest sources are S-EMBER (Wang et al., 2026c) with 109 items, AoTBench with 92, and ALLVB (Tan et al., 2025) with 45.

Protocols and scoring. The video protocol samples one frame per second, capped at 512 frames, and resizes frames to a short side of 224 pixels. Longer sequences are uniformly subsampled. Because the normalized store holds at most 1,024 frames, videos longer than 1,024 s are sampled below 1 fps. When memory or context limits reject a request, the runner halves the frame count and retries. It records the number of frames actually sent.

The blind protocol supplies the question and options without video. It averages accuracy over four option permutations per item using seed 42. Video gain is video accuracy minus blind accuracy. Every model receives the same introductory line, question, lettered choices and instruction to return an option letter. A shared letter-extraction scorer evaluates the response. Each model applies its released preprocessing to the shared frames.

Model implementations. We use the 8B variant when available, otherwise the largest smaller variant, namely VideoChat3-4B (Li et al., 2026d), LLaVA-OneVision-2-8B (An et al., 2026), Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025a), Molmo2-8B (Clark et al., 2026), VideoLLaMA3-7B (Zhang et al., 2025a), Video-XL-2 (Qin et al., 2025) at 7B, InternVideo2.5-Chat-8B (Wang et al., 2025d), and the standard Cambrian-S-7B (Yang et al., 2026a) checkpoint. Claude Opus 5 is the API reference. Qwen3-VL participates in pool construction and is excluded from ranking-agreement statistics.

Qwen3-VL uses vLLM 0.11 (Kwon et al., 2023). The remaining models use their released code. LLaVA-OneVision-2 and VideoChat3 run with transformers (Wolf et al., 2020) 5.16, Molmo2 and InternVL3.5 with 4.57, VideoLLaMA3, Video-XL-2, and Cambrian-S with 4.46, and InternVideo2.5 with 4.40. Each runs on one B200. InternVideo2.5 groups consecutive frames in fours and rounds its input count down to a multiple of four. VideoLLaMA3 falls back to 64 or 32 frames on 38% of items after rejecting other counts.

Comparison census. Table 39 reports a separate coverage census drawn from audit samples without attacker filtering. It takes up to eight items per benchmark, stratified by the benchmark's task labels, for 834 items across 107 benchmarks. The reported comparison gives reference accuracy 66.7% and blind accuracy 49.1% on this census. On the released Video-Index composition, reference accuracy is 56.8%, mean chance is 29.7%, and blind accuracy is 21.3%. Both video and blind accuracy are lower than on the coverage census.

Benchmark capability groups. The same labeler assigns groups to a stratified sample of 30 items per benchmark, totaling 3,124 items, and to all 720 sampled items of CG-Bench (Chen et al., 2025b), EgoPlan-Bench (Chen et al., 2025c), and PLM-VideoBench (Cho et al., 2025). Each benchmark takes its majority group. Perception covers 45 benchmarks, reasoning and knowledge 32, spatial and physical 20, and temporal 18. Only audited benchmarks contribute to the group totals.

Table 38: Video-Index accuracy (%) by duration, viewpoint, and presence/authenticity.

Model	Overall	Blind	Gain	Long video	First person	Authenticity
claude-opus-5	56.8	21.3	35.5	45.0	37.6	67.6
Molmo2-8B	19.0	4.8	14.2	9.9	10.2	26.8
InternVL3.5-8B	15.2	4.4	10.8	6.0	5.7	16.9
LLaVA-OneVision-2-8B	14.2	1.8	12.4	10.3	8.9	4.2
Video-XL-2	12.4	7.3	5.1	9.2	6.4	14.1
InternVideo2.5-8B	12.4	7.5	4.9	6.0	3.8	8.5
VideoLLaMA3-7B	12.4	5.2	7.1	5.7	6.4	7.0
Cambrian-S-7B	11.3	5.3	6.0	7.8	3.8	8.5
Qwen3-VL-8B†	9.5	1.4	8.1	8.5	6.4	5.6
VideoChat3-4B	9.4	3.8	5.7	4.6	3.8	15.5
Items	840	–	–	282	157	71

Table 39: Coverage-census accuracy and video gain (%) without attacker filtering (1/2).

Model	Overall	Blind	Gain
claude-opus-5	66.7	48.8	17.9
Molmo2-8B	58.8	40.3	18.4
InternVL3.5-8B	56.4	–	–
Qwen3-VL-8B†	55.1	39.1	16.1
LLaVA-OneVision-2-8B	54.5	36.5	18.1
Video-XL-2	53.8	39.8	14.0
VideoChat3-4B	51.8	38.9	12.8
Cambrian-S-7B	50.4	36.7	13.7
InternVideo2.5-8B	50.1	39.8	10.2
VideoLLaMA3-7B	49.5	16.0	33.5
Items	827	–	–

Table 40: Coverage-census accuracy and video gain (%) without attacker filtering (2/2).

Model	Perception	Temporal	Spatial	Reasoning
claude-opus-5	72.1	62.6	49.5	79.0
Molmo2-8B	66.0	51.6	47.8	63.6
InternVL3.5-8B	60.6	48.8	46.0	73.1
Qwen3-VL-8B†	62.5	49.7	44.6	58.0
LLaVA-OneVision-2-8B	59.9	56.1	44.0	54.5
Video-XL-2	62.5	49.7	41.3	55.1
VideoChat3-4B	57.4	47.7	38.6	59.1
Cambrian-S-7B	57.4	49.0	37.5	52.8
InternVideo2.5-8B	55.8	45.2	40.2	54.5
VideoLLaMA3-7B	55.4	45.2	39.7	52.8
Items	312	155	184	176

Constructing the coverage composition. The screening funnel leaves 62,142 multiple-choice items on 38,426 videos. No declared attacker solves them, every screening step is complete, and each video has a duration and scene label. The attackers use options only, question text, a single frame, 32 frames with the 2B model, or shuffled frames. We select 10,000 items to balance video durations, cover scene and question types, and retain difficulty under these attackers.

Duration groups and visual scenes. We use seven duration groups, namely <10 s, 10–30 s, 30–60 s, 1–3 min, 3–10 min, 10–30 min, and >30 min. The mean SigLIP2 embedding of 32 uniform frames approaches the global centroid for long videos with many scenes. We therefore detect visual near-duplicates separately within each duration group. Each threshold matches the random-pair false-match rate of cosine similarity at least 0.90 in the shortest group. In increasing duration order, the thresholds are 0.903, 0.853, 0.882, 0.946, 0.901, 0.901, and 0.975. Greedy leader clustering produces 19,263 scene clusters over 38,835 videos, as reported in Table 41.

Coverage and selection constraints. Selection cycles through duration groups, then the 19 scene types, then source benchmarks. Budget flows to strata that still contain admissible items. The initial caps allow one item per video, one video per scene cluster, five items per benchmark–question-template pair, and 5% of the budget per benchmark. A template replaces entities, numbers and nouns in the question with placeholders. Within each benchmark, we first prefer the least-templated item, then the item with the weakest strongest attacker. Attacker strength is the largest percentile rank among the five recorded margins.

A preliminary coverage pass selects one item per benchmark–question-type pair, covering 735 of 736 pairs. This pass ignores cluster and template caps to keep the rare question types of fully templated benchmarks. A duration–scene cell is scarce when its supply falls below an equal share of the budget. Such cells allow up to three items per video and lift cluster and template caps.

If a duration group still cannot fill its share, we raise the cluster cap to four and the template cap to ten, then lift its benchmark cap. These relaxations are needed for the 3–10 min, 10–30 min and >30 min groups. Only the 10–30 min group exceeds the benchmark cap, for S-EMBER (Wang et al., 2026c) and Video-MME-v2 (Fu et al., 2026a).

Selected composition. The result contains 10,000 items from 9,688 videos, 9,045 scene clusters, and 105 benchmarks. Every duration group includes all 19 scene types. The 230 videos with two or three selected items all belong to scarce cells. In increasing duration order, item counts are 1,432, 1,434, 1,432, 1,427, 1,427, 1,422, and 1,426. Median duration is 1.6 min.

The largest scene types are household daily life with 1,089 items, film and television narrative with 1,077, education and tutorials with 787, news and documentary with 766, and humor and advertising with 749. The smallest are industrial and construction scenes with 285 items, embodied robot manipulation with 270, and medical and first-aid scenes with 194.

A difficulty-first variant restricts selection to items whose worst attacker percentile is at most 0.7. It raises the share hard for all five attackers from 29% to 50%, but yields 1,015–1,884 items per duration group and places 12.8% in one benchmark. We retain the coverage-first composition, detailed in Tables 45 and 43.

Table 41: Visual-scene clusters by duration at matched false-match rates (1/2).

Duration	Videos	Items	Threshold	Scenes	Singletons
<10 s	9,445	11,043	0.903	6,003	5,243
10–30 s	11,592	14,943	0.853	4,823	3,356
30–60 s	5,146	7,457	0.882	2,254	1,610
1–3 min	3,122	5,313	0.946	1,787	1,396
3–10 min	5,028	9,178	0.901	2,013	1,333
10–30 min	2,248	4,376	0.901	1,030	681
>30 min	2,254	10,451	0.975	1,353	1,157
All	38,835	62,761	–	19,263	14,776

Table 42: Visual-scene clusters by duration at matched false-match rates (2/2).

Duration	Largest	Sharing	Scenes @0.90
<10 s	349	4,202 (44%)	5,943
10–30 s	248	8,236 (71%)	6,348
30–60 s	123	3,536 (69%)	2,537
1–3 min	117	1,726 (55%)	1,315
3–10 min	150	3,695 (73%)	1,999
10–30 min	79	1,567 (70%)	1,018
>30 min	136	1,097 (49%)	685
All	–	24,059 (62%)	19,845

Table 43: Coverage-pool composition by scene and duration; brackets give eligible counts (1/2).

Scene type	<10 s	10–30 s	30–60 s	1–3 min
household daily life	82 (829)	92 (5,424)	177 (3,209)	158 (1,507)
film tv narrative	81 (1,338)	91 (783)	81 (166)	144 (751)
education lecture tutorial	77 (199)	76 (429)	122 (346)	145 (347)
news documentary interview	71 (71)	68 (113)	66 (66)	91 (150)
humor memes advertising	80 (420)	76 (508)	148 (306)	141 (335)
sports exercise	82 (1,099)	75 (719)	97 (228)	106 (177)
cooking food	74 (196)	76 (467)	121 (401)	70 (108)
animals wildlife	74 (158)	78 (341)	97 (233)	119 (241)
ai generated synthetic	102 (2,606)	101 (1,550)	56 (665)	65 (534)
video games software ui	76 (221)	78 (283)	43 (109)	51 (51)
science experiment lab	75 (961)	82 (751)	98 (394)	72 (72)
music dance performance	76 (383)	78 (455)	50 (76)	46 (77)
other	75 (436)	74 (732)	51 (311)	34 (102)
traffic driving accidents	76 (674)	78 (906)	45 (199)	15 (15)
surveillance security anomaly	47 (202)	62 (196)	53 (55)	62 (68)
aerial drone urban navigation	77 (354)	66 (480)	18 (149)	30 (347)
industrial construction manufacturing	70 (440)	71 (172)	28 (78)	44 (44)
embodied robot manipulation	73 (311)	75 (181)	64 (117)	16 (16)
medical clinical first aid	64 (144)	37 (174)	17 (127)	18 (254)
All	1,432	1,434	1,432	1,427

Table 44: Coverage-pool composition by scene and duration; brackets give eligible counts (2/2).

Scene type	3–10 min	10–30 min	>30 min	Total
household daily life	131 (3,801)	366 (1,775)	83 (629)	1,089 (17,174)
film tv narrative	135 (1,331)	172 (479)	373 (6,545)	1,077 (11,393)
education lecture tutorial	131 (589)	175 (331)	61 (291)	787 (2,532)
news documentary interview	126 (344)	88 (198)	256 (769)	766 (1,711)
humor memes advertising	97 (224)	80 (192)	127 (360)	749 (2,345)
sports exercise	103 (326)	93 (222)	123 (447)	679 (3,218)
cooking food	85 (376)	91 (394)	55 (324)	572 (2,266)
animals wildlife	96 (559)	59 (152)	40 (100)	563 (1,784)
ai generated synthetic	64 (326)	40 (41)	39 (170)	467 (5,892)
video games software ui	59 (164)	65 (181)	41 (201)	413 (1,210)
science experiment lab	30 (478)	26 (134)	23 (23)	406 (2,813)
music dance performance	48 (127)	46 (51)	55 (266)	399 (1,435)
other	26 (152)	28 (118)	63 (73)	351 (1,924)

Table 44 (continued)

Scene type	3–10 min	10–30 min	>30 min	Total
traffic driving accidents	62 (64)	27 (27)	30 (37)	333 (1,922)
surveillance security anomaly	61 (68)	10 (13)	9 (21)	304 (623)
aerial drone urban navigation	66 (71)	21 (21)	18 (19)	296 (1,441)
industrial construction manufacturing	26 (95)	26 (38)	20 (160)	285 (1,027)
embodied robot manipulation	32 (34)	6 (6)	4 (10)	270 (675)
medical clinical first aid	49 (49)	3 (3)	6 (6)	194 (757)
All	1,427	1,422	1,426	10,000

Table 45: Benchmark contributions to the coverage pool and Video-Index, with attack results (1/2).

Benchmark	Class	Cand.	Remain.	Coverage set	Share (%)
ALLVB	Perception	24,955	5,008	139	1.4
Perception Test	Perception	17,803	3,145	101	1.0
VITATECS	Perception	11,203	623	7	0.1
HLV-1K	Perception	10,472	1,494	421	4.2
NExT-QA	Perception	8,550	213	84	0.8
FAVOR-Bench	Perception	7,985	2,162	77	0.8
ExpVid	Perception	5,521	1,768	177	1.8
VF-Eval	Perception	4,943	824	139	1.4
K9-Bench	Perception	4,687	1,086	305	3.0
ViMUL-Bench	Perception	4,483	1,046	244	2.4
UCF101-AD	Perception	4,224	7	5	0.1
VELOCITI	Perception	4,131	176	16	0.2
TemporalBench	Perception	3,748	630	179	1.8
MVBench	Perception	3,569	563	68	0.7
SITE-Bench	Perception	3,255	629	184	1.8
TaskMeAnything-v1	Perception	3,194	1,180	51	0.5
CAViAR	Perception	2,996	140	10	0.1
ST-Evidence	Perception	2,694	327	62	0.6
StreamingBench	Perception	2,380	419	203	2.0
TOC-Bench	Perception	2,211	1,030	97	1.0
MLVU	Perception	2,112	459	111	1.1
ScaleLong	Perception	1,707	381	161	1.6
LVBench	Perception	1,527	584	70	0.7
Colon-Bench	Perception	1,485	460	45	0.5
MME-VideoOCR	Perception	1,416	356	184	1.8
LongVideoBench	Perception	1,326	331	221	2.2
VideoEval-Pro	Perception	1,268	94	51	0.5
iSafetyBench	Perception	1,204	431	13	0.1
V2P-Bench	Perception	1,164	295	161	1.6
Video-LevelGauge	Perception	1,162	117	74	0.7
Causal2Needles	Perception	747	175	26	0.3
HERBench	Perception	628	156	48	0.5
Q-Bench-Video	Perception	622	84	69	0.7
TimeBlind	Perception	583	122	33	0.3
KFS-Bench	Perception	574	17	13	0.1
Artifact-Bench	Perception	565	30	5	0.1
VANE-Bench	Perception	556	64	39	0.4
NaturalBench-Video	Perception	458	90	38	0.4
GasVideo-1000	Perception	376	46	32	0.3
VSTAT	Perception	343	127	32	0.3
VNBench	Perception	320	95	37	0.4
FineBench	Perception	164	15	4	0.0
CG-Bench	Perception	–	–	–	–
PLM-VideoBench	Perception	–	–	–	0.0

Table 45 (continued)

Benchmark	Class	Cand.	Remain.	Coverage set	Share (%)
PerceptionComp	Perception	–	–	–	0.0
S-EMBER	Temporal	9,438	5,557	505	5.0
AoTBench	Temporal	5,294	830	262	2.6
MotionBench	Temporal	3,971	1,042	219	2.2
SLVMBench	Temporal	2,247	596	257	2.6
TVBench	Temporal	2,037	507	35	0.3
VidPair-Halluc	Temporal	1,907	388	37	0.4
TempCompass	Temporal	1,522	246	75	0.8
E.T. Bench	Temporal	1,500	503	26	0.3
TimeScope	Temporal	1,496	386	96	1.0
TOMATO	Temporal	1,421	741	142	1.4
VSI-Super-Wild	Temporal	1,193	600	116	1.2
Vinoground	Temporal	998	201	5	0.1
HourVideo	Temporal	993	589	53	0.5
ForeSeaQA	Temporal	532	64	56	0.6
VCR-Bench	Temporal	510	125	81	0.8
LongTimeScope	Temporal	450	182	28	0.3
PhysicalAI	Temporal	2	–	–	0.0
LEMONADE	Temporal	–	–	–	0.0
VGenST-Bench	Spatial / physical	9,470	1,374	216	2.2
VRU-Accident	Spatial / physical	5,988	834	30	0.3
SIS-Bench	Spatial / physical	4,052	715	89	0.9
UrbanVideo-Bench	Spatial / physical	3,917	1,396	85	0.8
ReVSI	Spatial / physical	3,381	1,126	51	0.5
CMU	Spatial / physical	3,228	485	38	0.4
M3-Verse	Spatial / physical	2,716	974	24	0.2
OSI-Bench	Spatial / physical	2,516	1,125	22	0.2
STI-Bench	Spatial / physical	1,885	820	128	1.3
DSI-Bench	Spatial / physical	1,695	414	26	0.3
SAW-Bench	Spatial / physical	1,561	676	28	0.3
ContPhy	Spatial / physical	1,379	102	11	0.1
DSR-Bench	Spatial / physical	1,334	703	246	2.5
OVO-S-Bench	Spatial / physical	1,172	348	125	1.2
MMSI-Video-Bench	Spatial / physical	1,085	495	144	1.4
IntPhys 2	Spatial / physical	1,064	145	6	0.1
VSI-Bench	Spatial / physical	694	125	21	0.2
VURB	Spatial / physical	407	39	17	0.2
SpatialTree	Spatial / physical	351	50	9	0.1
CineTechBench	Spatial / physical	127	36	5	0.1
GUI-World	Reasoning / knowledge	3,635	326	79	0.8
ExAct	Reasoning / knowledge	3,521	1,131	12	0.1
Video-Holmes	Reasoning / knowledge	3,384	1,038	69	0.7
COVER	Reasoning / knowledge	3,323	284	108	1.1
Video-MME-v2	Reasoning / knowledge	2,910	1,662	501	5.0
Video-MME	Reasoning / knowledge	2,698	530	308	3.1
VideoOdyssey	Reasoning / knowledge	2,447	1,091	105	1.1
VKnowU	Reasoning / knowledge	1,583	243	124	1.2
SFI-Bench	Reasoning / knowledge	1,530	366	20	0.2
MMWorld	Reasoning / knowledge	1,398	202	90	0.9
VideoEspresso	Reasoning / knowledge	1,359	396	116	1.2
VideoKR	Reasoning / knowledge	1,293	118	59	0.6
MMR-V	Reasoning / knowledge	1,218	536	156	1.6
VideoAds	Reasoning / knowledge	1,099	323	101	1.0
ReactHuman	Reasoning / knowledge	1,054	408	17	0.2
SmartHome-Bench	Reasoning / knowledge	1,023	77	18	0.2
Video-TT	Reasoning / knowledge	1,000	363	172	1.7
v-HUB	Reasoning / knowledge	951	219	74	0.7
SciVideoBench	Reasoning / knowledge	943	507	33	0.3
ReXTime	Reasoning / knowledge	918	108	78	0.8

Table 45 (continued)

Benchmark	Class	Cand.	Remain.	Coverage set	Share (%)
Video-MMMU	Reasoning / knowledge	873	267	95	0.9
LongVideo-Reason	Reasoning / knowledge	827	89	87	0.9
MMVU	Reasoning / knowledge	613	112	89	0.9
EgoSchema	Reasoning / knowledge	500	95	60	0.6
VideoMathQA	Reasoning / knowledge	419	110	76	0.8
VAU-Bench	Reasoning / knowledge	251	2	2	0.0
PAI-Bench-U	Reasoning / knowledge	2	1	1	0.0
EgoPlan-Bench	Reasoning / knowledge	–	–	–	0.0
MORSE-500	Reasoning / knowledge	–	–	–	0.0
Neptune	Reasoning / knowledge	–	–	–	0.0
SoMi-ToM	Reasoning / knowledge	–	–	–	0.0
VideoSIAH-Eval	Reasoning / knowledge	–	–	–	0.0

Table 46: Benchmark contributions to the coverage pool and Video-Index, with attack results (2/2).

Benchmark	Video-Index	ε_{opt}	$\varepsilon_{\text{text}}$	$\varepsilon_{\text{pool}}$	$\varepsilon_{\text{frame}}$	$\varepsilon_{\text{order}}$	Break
ALLVB	45	40	53	53	53	53	text
Perception Test	19	13	14	14	32	35	survives
VITATECS	24	26	29	29	39	41	frame
HLV-1K	7	33	42	42	45	45	text
NExT-QA	2	9	39	39	58	64	frame
FAVOR-Bench	19	26	34	34	34	34	text
ExpVid	1	25	28	28	40	40	frame
VF-Eval	18	52	58	58	58	58	text
K9-Bench	21	41	46	46	47	47	survives
ViMUL-Bench	7	16	20	20	39	44	survives
UCF101-AD	–	4	4	75	75	75	pool
VELOCITI	11	14	15	15	38	38	survives
TemporalBench	12	26	26	26	28	28	option
MVBench	1	8	18	18	34	37	order
SITE-Bench	2	17	24	24	34	39	order
TaskMeAnything-v1	1	4	7	18	19	40	order
CAViAR	–	23	23	29	41	41	frame
ST-Evidence	3	13	19	19	29	29	survives
StreamingBench	9	25	31	31	35	37	frame
TOC-Bench	38	4	4	8	15	15	frame
MLVU	6	22	34	34	41	41	frame
ScaleLong	4	21	24	24	33	33	survives
LVBench	8	8	12	12	22	22	frame
Colon-Bench	1	30	34	34	44	44	survives
MME-VideoOCR	5	14	18	18	33	42	survives
LongVideoBench	2	21	31	31	43	43	frame
VideoEval-Pro	3	7	10	10	23	23	frame
iSafetyBench	–	9	9	9	30	34	order
V2P-Bench	2	16	16	16	30	30	frame
Video-LevelGauge	2	14	28	28	43	64	order
Causal2Needles	–	21	32	32	35	35	frame
HERBench	–	16	28	28	34	34	frame
Q-Bench-Video	–	15	21	37	37	37	survives
TimeBlind	9	6	6	6	13	13	survives
KFS-Bench	–	19	30	30	39	39	survives
Artifact-Bench	–	-12	-9	30	30	30	option
VANE-Bench	1	51	52	53	62	62	frame
NaturalBench-Video	7	3	3	3	10	10	survives
GasVideo-1000	–	10	17	17	32	57	order
VSTAT	–	10	15	15	18	18	frame

Table 46 (continued)

Benchmark	Video-Index	ε_{opt}	$\varepsilon_{\text{text}}$	$\varepsilon_{\text{pool}}$	$\varepsilon_{\text{frame}}$	$\varepsilon_{\text{order}}$	Break
VNBench	–	2	2	3	27	27	survives
FineBench	–	36	36	60	60	60	option
CG-Bench	–	30	33	33	38	–	frame
PLM-VideoBench	–	2	22	22	32	–	frame
PerceptionComp	–	–	-13	-13	-11	-5	text
S-EMBER	109	11	11	11	11	11	survives
AoTBench	92	2	2	3	23	23	survives
MotionBench	11	21	25	25	31	31	survives
SLVMBench	21	30	43	43	53	53	survives
TVBench	3	2	7	10	16	16	survives
VidPair-Halluc	2	3	3	3	8	9	frame
TempCompass	13	15	15	18	18	18	option
E.T. Bench	2	3	5	6	25	25	survives
TimeScope	–	9	24	72	72	72	pool
TOMATO	4	3	3	12	12	12	survives
VSI-Super-Wild	–	4	5	5	5	5	option
Vinoground	18	1	1	1	27	27	survives
HourVideo	2	3	3	3	3	3	option
ForeSeaQA	1	44	46	46	46	46	option
VCR-Bench	1	18	21	21	27	27	survives
LongTimeScope	–	8	17	76	76	76	option
PhysicalAI	–	7	9	13	23	–	survives
LEMONADE	–	15	17	17	17	17	option
VGenST-Bench	6	12	19	28	44	44	survives
VRU-Accident	11	30	30	50	50	50	pool
SIS-Bench	5	16	25	25	37	37	survives
UrbanVideo-Bench	34	21	21	21	28	28	survives
ReVSI	2	3	13	13	13	23	order
CMU	33	17	22	22	34	34	survives
M3-Verse	1	2	3	10	10	15	order
OSI-Bench	1	5	14	14	14	17	survives
STI-Bench	–	22	27	27	29	29	text
DSI-Bench	6	27	27	29	30	42	option
SAW-Bench	3	4	4	36	36	36	pool
ContPhy	–	3	3	14	16	16	survives
DSR-Bench	1	4	4	8	8	19	order
OVO-S-Bench	2	13	13	16	16	16	option
MMSI-Video-Bench	1	3	6	6	6	7	option
IntPhys 2	–	2	2	2	2	8	frame
VSI-Bench	–	7	9	9	23	24	survives
VURB	–	22	22	29	29	38	order
SpatialTree	–	54	54	54	54	54	option
CineTechBench	–	6	15	15	32	32	frame
GUI-World	6	36	56	56	61	61	survives
ExAct	1	60	62	62	62	62	option
Video-Holmes	21	26	38	38	41	41	survives
COVER	–	16	32	32	35	37	text
Video-MME-v2	28	14	18	18	24	36	frame
Video-MME	9	21	31	31	40	47	order
VideoOdyssey	26	16	20	20	24	24	option
VKnowU	6	23	26	26	36	36	frame
SFI-Bench	3	23	37	37	37	37	survives
MMWorld	6	36	52	52	55	57	text
VideoEspresso	7	49	51	51	54	54	text
VideoKR	–	20	23	23	35	50	order
MMR-V	1	39	47	47	48	48	text
VideoAds	31	15	26	26	51	51	frame
ReactHuman	–	4	4	4	12	19	frame
SmartHome-Bench	1	13	13	13	21	21	survives

Table 46 (continued)

Benchmark	Video-Index	ε_{opt}	$\varepsilon_{\text{text}}$	$\varepsilon_{\text{pool}}$	$\varepsilon_{\text{frame}}$	$\varepsilon_{\text{order}}$	Break
Video-TT	9	5	12	32	32	33	pool
v-HUB	–	5	5	5	48	50	frame
SciVideoBench	–	40	44	44	48	48	survives
ReXTime	4	59	60	60	66	66	frame
Video-MMMU	1	30	45	45	45	59	order
LongVideo-Reason	1	35	49	49	50	50	text
MMVU	1	29	38	38	58	58	frame
EgoSchema	1	37	38	38	51	57	order
VideoMathQA	2	4	12	25	25	25	option
VAU-Bench	–	63	71	71	71	71	text
PAI-Bench-U	–	6	9	10	25	25	survives
EgoPlan-Bench	–	3	8	8	8	–	option
MORSE-500	–	7	7	15	15	15	survives
Neptune	–	3	52	52	52	52	survives
SoMi-ToM	–	7	17	20	37	61	order
VideoSIAH-Eval	–	–	-18	-18	-13	-13	text

We assess five properties of a composition. Validity combines residual attack scores, blind accuracy, and video gain. Reliability uses frame-offset and option-permutation noise, 1,000 item-bootstrap resamples, and the ratio of score dispersion to the noise floor. Ordering checks pairs with an expected direction, including model sizes, frame budgets, and video versus blind inputs. We count correctly ordered pairs with non-overlapping 95% intervals and report separation per hundred items. Agreement uses Kendall τ against mean chance-corrected accuracy over 3 to 8 stratified items from each of 106 census benchmarks. Each source is excluded from the census when its own agreement is evaluated. We repeat the ranking comparison using video gain to distinguish dependence on video from shortcut-driven accuracy. Hardness combines reference accuracy with the share of failures attributed to labels. Models used in screening or gating are excluded from ordering and agreement statistics.

Table 47: Selection at 840 items: coverage, shortcut exposure, video gain, and rank agreement.

Selection	n	Cells	Exploit. (%)	Gain (%)	SNR	Pairs (%)	τ_{acc}	τ_{gain}
Video-Index (meta-benchmark)	840	354	13.1	10.8	10.8	77	0.33	0.56
Random draw from the pool (mean of 5)	840	374	19.3	11.7	9.9	86	0.36	0.43
Top benchmarks by breaking level	840	234	19.1	12.1	9.1	77	0.22	0.50
Diversity-aware draw, no gate	840	578	19.4	11.3	11.2	77	0.28	0.59
Coverage census (≤ 8 per benchmark)	834	–	–	15.5	2.6	90	0.94	1.00
Median source benchmark	54	–	19.0	9.4	3.7	38	0.30	0.22

Video-Index separates all ten video-versus-blind pairs and 77% of the known-order pairs, the same share as selection by breaking level, against 86% for the random draws. Table 47 reports Kendall $\tau = 0.56$ with the census ranking by video gain, comparable to the alternatives. Agreement by accuracy is 0.33. The gate provides the clearest distinction. Strongest-attacker accuracy is 13%, versus 19% without gating. As shown in Figure U3, gain-based agreement rises with the budget, from $\tau = 0.34$ at 105 items to 0.49 at 420 and 0.56 at 840.

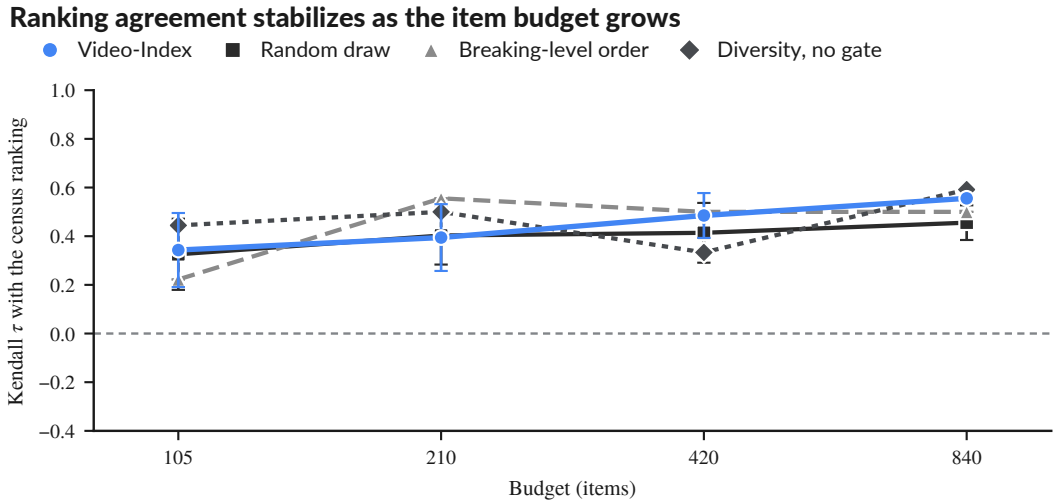


Figure U3: Kendall τ agreement of model ranks with the census across evaluation budgets.

V AGENT BEHAVIOR ON VIDEO-INDEX

Claude Opus 5 (Anthropic, 2026b) is evaluated through the same type of tool-enabled CLI protocol, with maximum effort, model-directed sampling, and failures counted as incorrect. It scores 70.6% and 24.9% blind over four option permutations. The fixed 1 fps / 512-frame protocol gives 56.8% with video and 21.3% blind.

Figure 12 takes accuracy from the leaderboard and effort from the session logs. Shell calls, elapsed time, and image counts are medians per session. The horizontal runtime intervals in Figure 12 span the first and third quartiles of recorded session times. Runtime quartiles use the sessions with a recorded time. Each operation bar measures the percentage of that agent's logged answers using the operation. The full background track represents 100%. The four percentages are independent because operations can co-occur. Visual operations are counted once per session. Regular sampling includes fixed-rate, fixed-count, and every-nth-frame extraction. Seeking selects timestamps. Tiling assembles frame grids. Cropping or zooming isolates details. Mixed strategies can contribute to multiple operations. Images produced include decoded frames, crops, and contact sheets, whether or not the model viewed them. Image views are counted where the logs record them. The duration panel uses the exported per-agent medians. Tables 48 and 49 report the remaining tool and sampling details.

Sampling choices differ substantially. Fable decodes every stored frame in 72.2% of sessions. Gemini Pro uses an explicit frame-rate filter in 47.6%. Astra and Opus distribute their sessions across full decoding, periodic sampling, and mixed strategies. Contact sheets appear in 99–100% of sessions for Astra, Fable, and Opus, compared with 69% for Gemini Pro and 35% for Flash-Lite. Flash-Lite retains native extraction resolution in every logged session, while Astra downscales in 60.8%. These choices describe different inspection policies.

Among Astra, Fable, and Opus, pairwise answer agreement is 76.8–80.6%, while strategy agreement is 33.0–43.8%. Similar answers can therefore follow different inspection paths. Long videos remain difficult. Beyond ten minutes, accuracy is 67.1% for Astra, 50.9% for Fable, and 46.6% for Opus, down from 88.7%, 84.4%, and 86.2% below 30 seconds in the answer export. Table 50 gives the tool-log summaries by duration, including both Gemini agents. Inspection volume also varies with item difficulty. From the lowest to highest image-production quartile, accuracy changes from 85.4% to 74.1% for Astra, from 85.4% to 51.6% for Fable, and from 73.0% to 62.2% for Opus. These associations do not show that producing more images causes worse answers. Separate vendor usage records report median input/output tokens of 31k/0.9k for Astra and 128k/10k for Fable. Input counts include cache reads.

Table 48: Agent tool use on Video-Index, as medians per item unless indicated otherwise.

Measure	Astra	Fable 5.1	Gemini 3.1 Pro	Flash-Lite	Opus 5
Accuracy (%)	79.3	71.6	55.3	21.7	70.6
Shell commands per item	3	5	4	6	12
Sessions using ffmpeg (%)	50	100	100	100	100
Sessions using OpenCV (%)	54	16	26	96	48
Sessions using Pillow (%)	81	81	51	61	90
Sessions with a contact sheet (%)	99	100	69	35	99
Sessions with crops or zooms (%)	11	74	24	9	84
Images produced per item	8	79	39	62	25
Images viewed per item	–	12	6	6	13
Images produced per video second	0.10	2.19	1.63	2.00	0.73
Seconds per item	57	200	140	95	205
Output tokens per item	824	10239	–	–	13293

Table 49: Agent sampling strategies (% of sessions); videos store at most 2 fps and 1,024 frames.

Strategy	Astra	Fable 5.1	Gemini 3.1 Pro	Flash-Lite	Opus 5
Decode every stored frame (2 fps)	35.0	72.2	19.4	39.1	26.3
Every n-th frame	13.4	22.5	14.6	8.7	33.5
Explicit fps filter	6.5	0.7	47.6	8.7	6.2
Fixed number of frames	1.5	1.4	–	–	0.8
Seeks to chosen timestamps	7.6	0.5	–	–	1.6
Two or more of the above	36.0	2.7	18.4	43.5	31.4
No decoding command	–	–	–	–	0.1

Table 50: Agent sampling strategies, image counts, and accuracy by video duration.

Video duration	Agent	Strategy	Share (%)	Images	Acc. (%)
<30 s (335 items)	Astra	all	70	3	89
	Fable 5.1	all	79	34	85
	Gemini Pro	fps	52	20	68
	Flash-Lite	all	82	16	27
	Opus 5	nth	44	30	86
30-120 s (114 items)	Astra	mix	48	3	76
	Fable 5.1	all	77	114	67
	Gemini Pro	fps	54	63	64
	Flash-Lite	nth	100	62	100
	Opus 5	mix	44	24	66
120-600 s (141 items)	Astra	mix	88	7	74
	Fable 5.1	all	51	636	69
	Gemini Pro	fps	56	184	56
	Flash-Lite	mix	83	614	0
	Opus 5	mix	76	18	65
>600 s (164 items)	Astra	mix	52	53	67
	Fable 5.1	all	73	1060	51
	Gemini Pro	nth	40	156	15
	Flash-Lite	mix	100	900	20
	Opus 5	mix	38	24	47

W FULL TABLES

Table 51 summarizes capability groups and their audit statistics. Tables 52–53 collect benchmark-level results, ranked by the accuracy gap between 32-frame and blind inputs. In Table 52, the caption column uses Claude Opus 5 (Anthropic, 2026b) for every row. Lift is blind accuracy minus chance, where chance is the mean item chance over the multiple-choice items taken from the option list, the options embedded in the question, the declared option count, or an inferred count, in that order. The reasoning and protocol columns give the share of attributed errors in reasoning and knowledge and in data protocol. The two attribution protocols use different judge pairs, as Tables 30 and 32 list. Diversity is the Vendi score per item, and the membership column is the min-20% score under Qwen2.5-7B (Qwen Team, 2024), where a lower value marks more corpus exposure. Diversity, near-duplicate, and membership cells are blank where not computed.

Table 51: Capability groups: counts, video gain, protocol issues, and largest-gain sources.

Category	n	Median m	Median proto.	Highest-m member
Perception	45	16.0	6	VNBench
Temporal	18	16.7	9	Vinoground
Spatial / physical	20	14.0	2	VGenST-Bench
Reasoning / knowledge	32	15.0	22	v-HUB

Table 52: Audit results ranked by 32-frame accuracy minus blind accuracy (1/2).

#	Benchmark	Yr	Chance	Blind	1f	32f	Cap.	m
2	VNBench	2025	25	26	25	74	60	47.9
3	v-HUB	2025	20	17	52	64	71	47.4
4	SoMi-ToM	2025	33	51	47	93	85	42.7
5	VGenST-Bench	2026	27	45	61	88	81	42.3
6	MME-VideoOCR	2025	25	43	58	84	68	41.3
7	UCF101-AD	2026	9	7	36	47	49	39.3
8	EgoSchema	2023	20	39	50	77	80	37.6
9	Video-LevelGauge	2026	25	53	65	89	74	36.3
10	Vinoground	2024	50	46	56	82	79	36.2
11	ViMUL-Bench	2025	26	44	62	80	75	36.2
12	Video-TT	2025	25	28	27	61	45	32.4
13	CAViAR	2026	32	45	73	78	75	32.3
14	VideoKR	2026	14	23	30	55	57	32.1
15	MORSE-500	2025	20	10	1	41	24	30.9
16	VideoAds	2025	25	51	57	81	75	30.3
17	NaturalBench-Video	2025	50	51	55	81	67	30.0
18	MVBench	2023	29	39	42	69	66	29.9
19	TaskMeAnything-v1	2024	25	32	41	62	50	29.7
20	GasVideo-1000	2026	30	46	48	75	73	29.2
21	AoTBench	2025	47	47	46	77	72	29.1
22	VELOCITI	2025	50	65	75	94	93	28.3
23	Q-Bench-Video	2024	27	44	59	73	46	28.3
24	VRU-Accident	2025	25	48	67	76	75	27.7
25	Perception Test	2023	33	47	54	75	71	27.7
26	iSafetyBench	2025	6	15	36	42	43	27.4
27	MotionBench	2025	25	38	48	65	64	26.5
28	PAI-Bench-U	–	25	34	50	61	56	26.5
29	E.T. Bench	2024	23	29	33	55	52	26.0
30	TVBench	2024	33	37	30	63	56	25.5
31	VKnowU	2025	36	46	42	71	77	25.5
32	VSI-Bench	2024	29	36	17	61	45	24.5
33	VITATECS	2023	50	68	82	92	89	23.6
34	ContPhy	2024	31	33	47	56	34	22.7

Table 52 (continued)

#	Benchmark	Yr	Chance	Blind	1f	32f	Cap.	m
35	VURB	2026	27	40	37	62	58	22.6
36	SITE-Bench	2025	26	43	38	65	65	22.2
37	Video-MMMU	2025	12	48	44	70	70	22.0
38	SIS-Bench	–	25	50	58	72	69	22.0
39	TimeBlind	2026	50	49	25	71	66	21.7
40	TOMATO	2024	20	22	13	43	30	21.4
41	VCR-Bench	2025	24	45	47	66	58	21.2
42	ReVSI	2026	27	31	9	52	40	20.7
43	ST-Evidence	2026	47	62	53	82	81	20.5
44	TempCompass	2024	27	22	28	42	40	20.2
45	CineTechBench	2025	25	40	50	60	55	20.0
46	MMVU	2025	20	57	72	77	84	20.0
47	Colon-Bench	2026	20	54	64	74	62	20.0
48	OSI-Bench	–	25	29	30	48	42	19.2
49	ReactHuman	2026	33	31	40	50	51	18.9
50	Video-MME	2024	25	52	41	70	71	18.2
51	Video-MME-v2	2026	13	19	17	36	39	17.7
52	LVBench	2024	25	29	29	46	47	16.9
53	SLVMBench	2026	20	63	65	80	73	16.7
54	TimeScope	2025	17	36	16	53	58	16.6
55	MLVU	2024	25	48	33	65	69	16.1
56	KFS-Bench	2026	21	51	52	67	62	16.0
57	MMWorld	2024	26	59	61	74	81	15.8
58	VidPair-Halluc	2026	41	27	42	43	42	15.8
59	V2P-Bench	2025	25	37	43	53	53	15.7
60	StreamingBench	2024	25	46	35	62	62	15.5
61	ScaleLong	2025	25	49	53	64	60	15.3
62	SmartHome-Bench	2025	50	63	71	78	71	15.0
63	TOC-Bench	2026	32	35	41	50	47	14.6
64	SFI-Bench	2026	25	55	21	69	61	14.3
65	M3-Verse	2025	24	26	34	40	34	14.0
66	Video-Holmes	2025	17	54	58	68	63	13.3
67	PLM-VideoBench	2025	50	72	82	85	80	13.2
68	UrbanVideo-Bench	2025	21	41	41	54	44	12.7
69	ExpVid	2025	24	52	65	65	63	12.4
70	SpatialTree	2025	17	51	23	62	59	11.4
71	GUI-World	2025	25	81	83	92	88	11.0
72	VANE-Bench	2025	25	77	87	88	82	11.0
73	SciVideoBench	2025	10	54	52	65	62	11.0
74	LongVideoBench	2024	21	51	26	60	70	9.7
75	VideoEval-Pro	2025	25	35	25	45	46	9.6
76	SAW-Bench	2026	26	29	17	38	32	9.4
77	K9-Bench	2026	20	66	64	75	69	9.3
78	CG-Bench	2024	15	48	52	57	48	9.2
79	VideoSIAH-Eval	2025	25	4	4	13	11	8.9
80	FAVOR-Bench	2025	20	47	28	55	54	8.6
81	MMR-V	2025	10	47	40	56	58	8.6
82	COVER	2025	25	51	45	59	60	8.1
83	S-EMBER	2026	20	29	22	37	24	7.7
84	HERBench	2026	20	48	50	56	55	7.7
85	ReXTime	2024	25	84	88	92	90	7.5
86	PerceptionComp	2026	20	7	5	13	9	6.8
87	OVO-S-Bench	2026	30	40	42	47	48	6.6
88	Causal2Needles	2025	25	57	59	64	60	6.3
89	IntPhys 2	2025	50	52	50	57	55	5.4
90	ExAct	2025	20	64	61	69	79	5.2
91	TemporalBench	2024	50	72	68	77	76	5.1
92	VideoEspresso	2025	25	75	79	80	77	4.6
93	VSI-Super-Wild	2026	25	29	13	32	31	3.5

Table 52 (continued)

#	Benchmark	Yr	Chance	Blind	1f	32f	Cap.	m
94	Artifact-Bench	2026	17	6	8	9	11	3.2
95	ForeSeaQA	2026	25	71	71	73	69	2.5
96	HLV-1K	2025	25	67	66	69	70	1.9
97	STI-Bench	2025	20	47	49	49	48	1.7
98	DSI-Bench	2025	25	52	44	53	60	1.2
99	LongVideo-Reason	2025	27	76	73	77	79	1.0
100	VF-Eval	2025	25	83	81	84	82	0.7
102	DSR-Bench	–	25	24	26	24	34	-0.1
103	ALLVB	2025	20	73	70	73	71	-0.7
104	FineBench	2026	25	63	57	62	63	-1.4
105	HourVideo	2024	20	23	10	21	24	-1.5
106	VAU-Bench	2025	25	96	95	94	95	-1.6
107	LEMONADE	2025	25	42	32	40	38	-1.7
108	MMSI-Video-Bench	2025	24	30	17	28	35	-1.8
109	EgoPlan-Bench	2023	25	33	28	29	30	-3.7
110	VideoMathQA	2025	44	56	26	52	69	-4.5
111	VideoOdyssey	2026	25	44	23	35	49	-9.6
112	LongTimeScope	–	17	34	4	16	27	-17.9
113	NExT-QA	2021	20	–	–	–	79	–
114	Neptune	2024	20	72	71	81	–	–
115	VSTAT	2026	26	–	–	–	50	–

Table 53: Audit results ranked by 32-frame accuracy minus blind accuracy (2/2).

#	Benchmark	Lift	Reason.	Proto.	Div.	NearDup	mink
2	VNBench	1.0	87	4	0.12	58	7.0
3	v-HUB	-3.1	39	50	0.21	22	9.8
4	SoMi-ToM	17.3	67	0	0.05	100	9.7
5	VGenST-Bench	18.7	67	0	0.25	28	8.0
6	MME-VideoOCR	17.7	69	3	0.35	10	7.3
7	UCF101-AD	-1.8	10	21	0.06	92	6.4
8	EgoSchema	19.3	77	22	0.12	66	6.7
9	Video-LevelGauge	27.7	62	0	0.34	22	8.0
10	Vinoground	-4.0	72	0	0.06	81	7.5
11	ViMUL-Bench	18.3	52	49	0.28	10	7.3
12	Video-TT	3.3	58	14	0.31	13	7.7
13	CAViAR	13.2	69	38	0.02	100	6.2
14	VideoKR	9.1	82	52	0.25	26	8.0
15	MORSE-500	-9.8	29	16	0.02	100	7.4
16	VideoAds	26.0	63	23	0.20	27	6.3
17	NaturalBench-Video	0.7	89	0	0.22	67	8.3
18	MVBench	10.3	38	8	0.15	67	9.2
19	TaskMeAnything-v1	6.7	45	0	0.08	84	8.5
20	GasVideo-1000	16.1	67	71	0.27	39	11.0
21	AoTBench	0.7	71	0	0.16	35	8.8
22	VELOCITI	15.3	38	0	0.02	99	7.2
23	Q-Bench-Video	17.0	52	38	0.19	12	7.8
24	VRU-Accident	23.3	80	9	0.02	100	7.8
25	Perception Test	13.7	55	4	0.16	77	9.0
26	iSafetyBench	8.5	30	4	0.05	99	6.8
27	MotionBench	13.0	48	9	0.22	29	6.9
28	PAI-Bench-U	9.0	66	0	0.22	46	9.1
29	E.T. Bench	5.3	42	0	0.06	97	8.1
30	TVBench	4.0	44	13	0.09	90	8.2
31	VKnowU	9.9	76	29	0.23	42	6.9
32	VSI-Bench	7.5	20	8	0.07	98	7.9

Table 53 (continued)

#	Benchmark	Lift	Reason.	Proto.	Div.	NearDup	mink
33	VITATECS	18.5	67	14	0.06	79	7.1
34	ContPhy	2.3	44	0	0.05	100	8.6
35	VURB	13.2	50	42	0.22	61	8.0
36	SITE-Bench	17.3	39	9	0.21	47	9.1
37	Video-MMMU	36.6	83	38	0.30	23	6.5
38	SIS-Bench	25.0	33	0	0.06	85	9.2
39	TimeBlind	-0.8	53	6	0.19	80	7.9
40	TOMATO	2.1	42	10	0.15	72	9.4
41	VCR-Bench	21.1	60	2	0.34	9	8.9
42	ReVSI	4.3	21	12	0.09	94	8.2
43	ST-Evidence	14.2	59	15	0.24	35	7.9
44	TempCompass	-4.5	75	57	0.17	58	8.6
45	CineTechBench	15.2	45	0	0.07	88	8.4
46	MMVU	36.9	80	27	0.25	25	7.3
47	Colon-Bench	34.3	56	0	0.04	100	6.8
48	OSI-Bench	3.6	20	4	0.08	91	8.2
49	ReactHuman	-2.0	58	0	0.01	100	8.4
50	Video-MME	27.0	69	40	0.27	5	7.0
51	Video-MME-v2	6.0	73	28	0.31	15	6.5
52	LVBench	3.7	70	40	0.15	68	8.3
53	SLVMBench	43.0	79	7	0.20	51	6.3
54	TimeScope	19.3	70	82	0.07	100	9.7
55	MLVU	23.5	69	59	0.11	77	7.8
56	KFS-Bench	30.2	71	7	0.26	27	7.2
57	MMWorld	32.7	77	26	0.23	53	7.6
58	VidPair-Halluc	-14.2	83	31	0.14	75	9.0
59	V2P-Bench	11.9	67	49	0.19	25	9.7
60	StreamingBench	21.4	51	31	0.21	54	7.1
61	ScaleLong	23.7	74	10	0.27	28	6.5
62	SmartHome-Bench	13.0	67	29	0.02	99	5.9
63	TOC-Bench	2.9	69	1	0.16	42	8.7
64	SFI-Bench	29.7	58	3	0.14	59	7.6
65	M3-Verse	2.7	50	1	0.02	100	7.8
66	Video-Holmes	37.7	88	4	0.16	36	8.5
67	PLM-VideoBench	21.7	62	0	0.24	19	6.3
68	UrbanVideo-Bench	20.3	58	1	0.07	94	7.4
69	ExpVid	27.8	56	0	0.11	69	8.7
70	SpatialTree	33.3	33	16	0.08	79	7.4
71	GUI-World	56.0	72	5	0.27	12	7.0
72	VANE-Bench	52.3	62	0	0.16	72	6.9
73	SciVideoBench	44.0	95	2	0.08	76	7.8
74	LongVideoBench	29.3	68	59	0.30	15	8.2
75	VideoEval-Pro	10.7	79	47	0.30	17	7.4
76	SAW-Bench	2.9	39	0	0.04	97	8.8
77	K9-Bench	45.5	52	0	0.15	41	7.5
78	CG-Bench	32.9	48	8	–	–	6.4
79	VideoSIAH-Eval	-21.1	42	82	0.32	8	10.5
80	FAVOR-Bench	26.7	46	10	0.19	26	8.0
81	MMR-V	37.4	96	25	0.20	24	7.5
82	COVER	26.1	65	31	0.26	8	7.4
83	S-EMBER	9.4	45	0	0.22	12	6.5
84	HERBench	28.0	53	0	0.06	99	7.0
85	ReXTime	59.3	73	8	0.31	6	7.0
86	PerceptionComp	-13.3	33	17	0.17	67	8.7
87	OVO-S-Bench	9.9	74	2	0.08	91	7.3
88	Causal2Needles	32.3	60	0	0.14	42	9.0
89	IntPhys 2	1.7	17	19	0.02	100	10.6
90	ExAct	44.0	74	6	0.02	100	6.5
91	TemporalBench	21.8	63	5	0.14	62	6.7

Table 53 (continued)

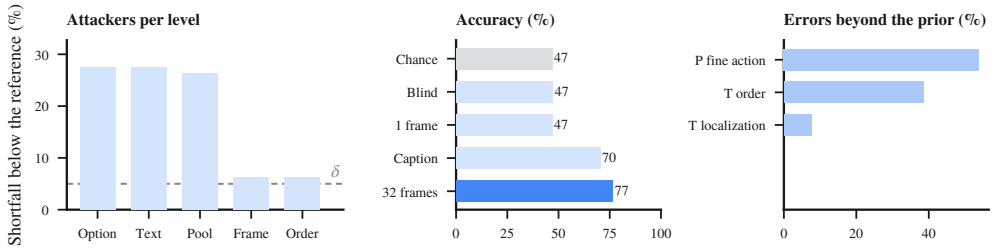
#	Benchmark	Lift	Reason.	Proto.	Div.	NearDup	mink
92	VideoEspresso	50.0	90	12	0.26	23	6.8
93	VSI-Super-Wild	4.0	55	15	0.04	97	7.6
94	Artifact-Bench	-11.0	47	2	0.11	49	10.4
95	ForeSeaQA	45.9	54	0	0.13	84	6.5
96	HLV-1K	42.0	57	4	0.16	46	6.7
97	STI-Bench	27.2	43	1	0.10	88	8.1
98	DSI-Bench	26.5	27	1	0.09	67	6.7
99	LongVideo-Reason	49.5	94	8	0.21	11	8.1
100	VF-Eval	58.3	61	0	0.18	30	7.6
102	DSR-Bench	-0.9	41	4	0.16	49	6.4
103	ALLVB	53.3	89	9	0.17	24	6.7
104	FineBench	38.3	77	0	0.04	100	7.5
105	HourVideo	2.7	62	28	0.14	74	7.0
106	VAU-Bench	71.0	43	0	0.12	70	6.3
107	LEMONADE	16.7	57	3	0.05	100	10.7
108	MMSI-Video-Bench	5.9	41	14	0.16	48	7.0
109	EgoPlan-Bench	8.0	68	12	0.04	97	9.0
110	VideoMathQA	11.7	71	43	0.20	63	7.3
111	VideoOdyssey	19.4	79	51	0.30	27	7.1
112	LongTimeScope	17.0	50	88	0.05	100	9.7
113	NExT-QA	–	–	–	0.26	7	9.1
114	Neptune	–	33	73	0.28	14	8.8
115	VSTAT	–	–	–	0.06	100	8.7

X REPORT CARDS

One card per audited benchmark, the page released with the pool. Each card gives attack gaps, breaking level, reference and shortcut accuracies, dominant error categories, effective size, near-duplicate share, video dependence, protocol-error share, and pool contribution. The cards regenerate from the same exported tables as the paper.

AoTBench

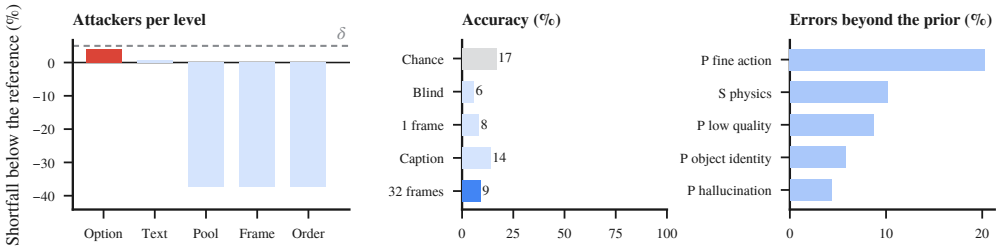
released 2025, chance 47%, reference 77% at 32 frames, survives every level



Effective size 32 of 200, near-duplicates 35%, video dependence +29%, protocol issues 0% of errors, coverage pool 262, Video-Index 92

Artifact-Bench

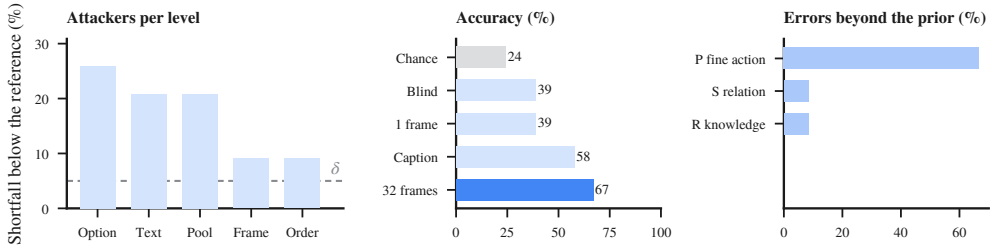
released 2026, chance 17%, reference 9% at 32 frames, breaks at the option level



Effective size 22 of 200, near-duplicates 49%, video dependence +3%, protocol issues 2% of errors, coverage pool 5, Video-Index 0

CMU

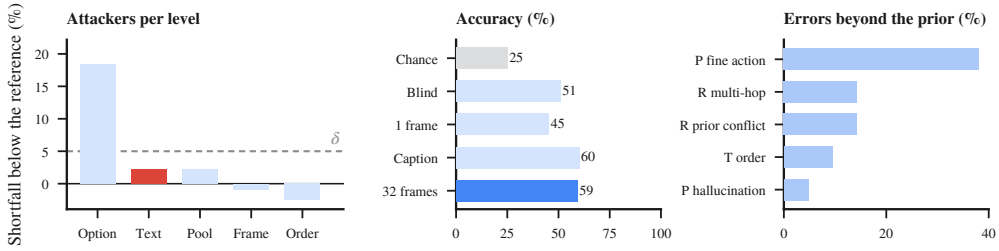
released 2026, chance 24%, reference 67% at 32 frames, survives every level



Effective size 33 of 200, near-duplicates 28%, video dependence +28%, protocol issues n/a of errors, coverage pool 38, Video-Index 33

COVER

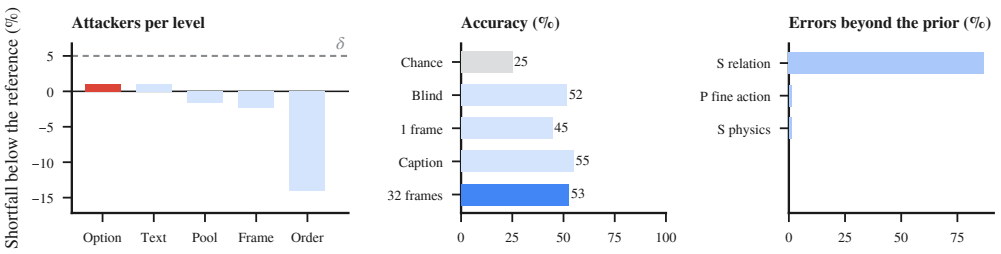
released 2025, chance 25%, reference 59% at 32 frames, breaks at the text level



Effective size 52 of 200, near-duplicates 8%, video dependence +8%, protocol issues 31% of errors, coverage pool 108, Video-Index 0

DSI-Bench

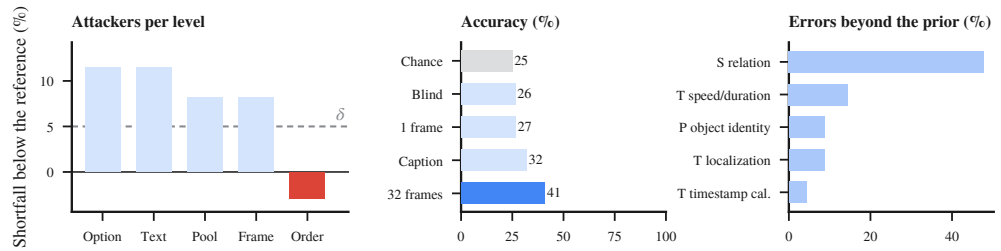
released 2025, chance 25%, reference 53% at 32 frames, breaks at the option level



Effective size 18 of 200, near-duplicates 67%, video dependence +1%, protocol issues 1% of errors, coverage pool 26, Video-Index 6

DSR-Bench

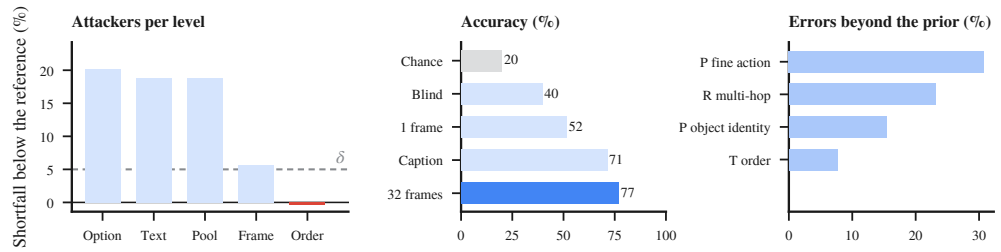
released 2025, chance 25%, reference 41% at 128 frames, breaks at the order level



Effective size 31 of 200, near-duplicates 49%, video dependence +14%, protocol issues 4% of errors, coverage pool 246, Video-Index 1

EgoSchema

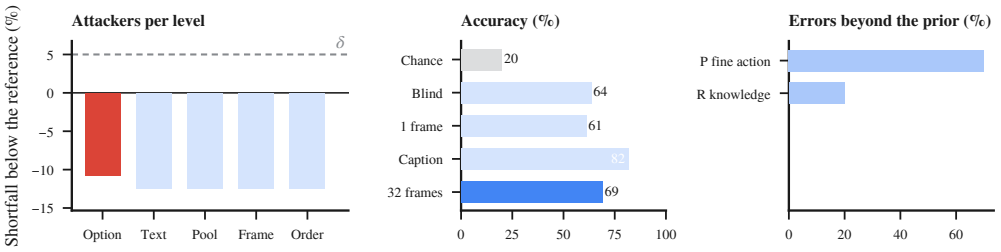
released 2023, chance 20%, reference 77% at 32 frames, breaks at the order level



Effective size 24 of 200, near-duplicates 66%, video dependence +37%, protocol issues 22% of errors, coverage pool 60, Video-Index 1

ExAct

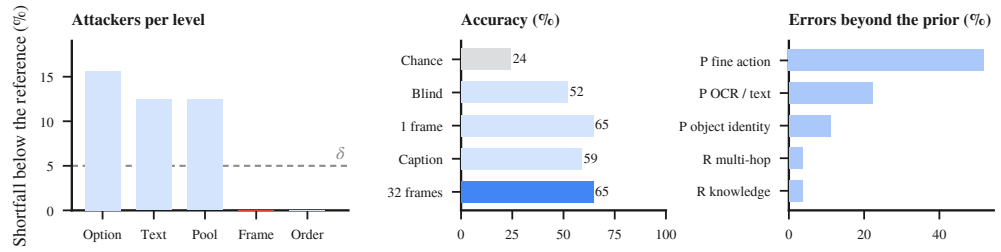
released 2025, chance 20%, reference 69% at 32 frames, breaks at the option level



Effective size 4 of 200, near-duplicates 100%, video dependence +6%, protocol issues 6% of errors, coverage pool 12, Video-Index 1

ExpVid

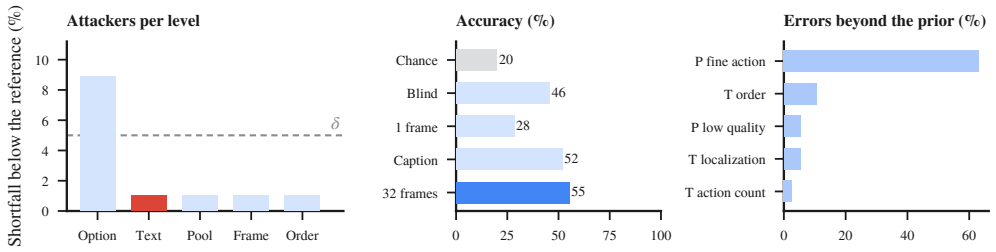
released 2025, chance 24%, reference 65% at 32 frames, breaks at the frame level



Effective size 22 of 200, near-duplicates 69%, video dependence +12%, protocol issues 0% of errors, coverage pool 177, Video-Index 1

FAVOR-Bench

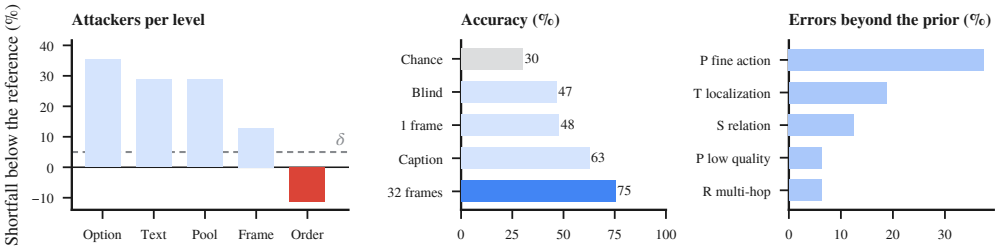
released 2025, chance 20%, reference 55% at 32 frames, breaks at the text level



Effective size 39 of 200, near-duplicates 26%, video dependence +10%, protocol issues 10% of errors, coverage pool 77, Video-Index 19

GasVideo-1000

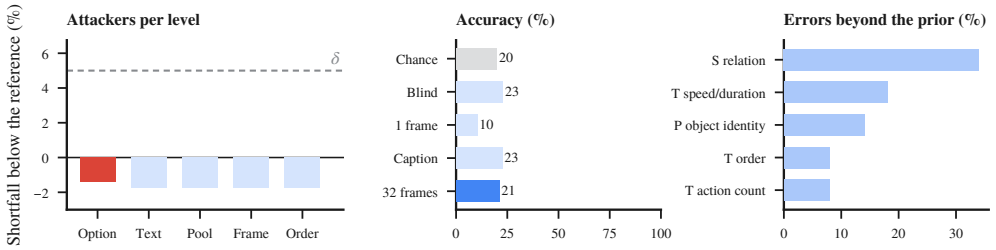
released 2026, chance 30%, reference 75% at 32 frames, breaks at the order level



Effective size 54 of 200, near-duplicates 39%, video dependence +29%, protocol issues 71% of errors, coverage pool 32, Video-Index 0

HourVideo

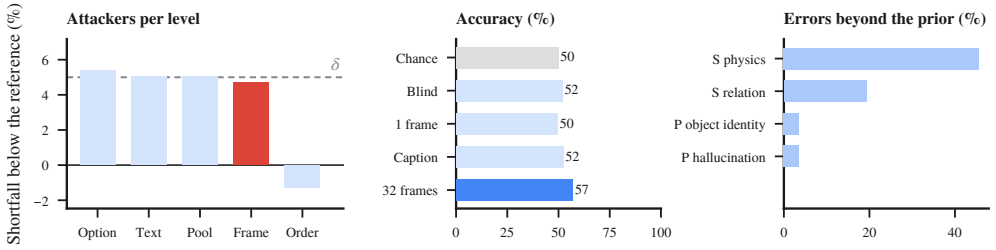
released 2024, chance 20%, reference 21% at 32 frames, breaks at the option level



Effective size 28 of 200, near-duplicates 74%, video dependence -2%, protocol issues 28% of errors, coverage pool 53, Video-Index 2

IntPhys 2

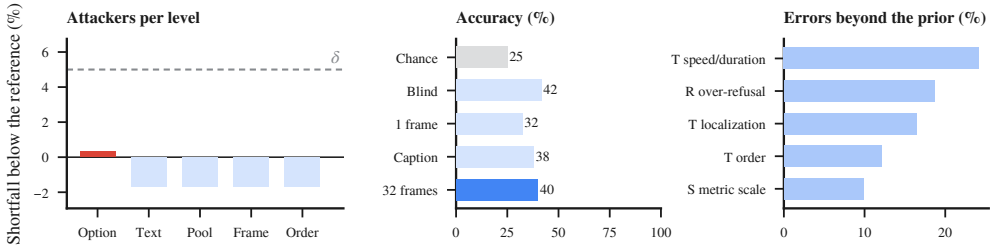
released 2025, chance 50%, reference 57% at 32 frames, breaks at the frame level



Effective size 5 of 200, near-duplicates 100%, video dependence +5%, protocol issues 19% of errors, coverage pool 6, Video-Index 0

LEMONADE

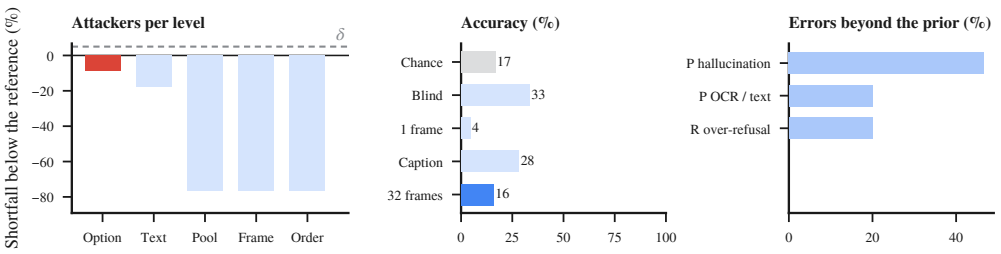
released 2025, chance 25%, reference 40% at 32 frames, breaks at the option level



Effective size 9 of 200, near-duplicates 100%, video dependence -2%, protocol issues 3% of errors, coverage pool 0, Video-Index 0

LongTimeScope

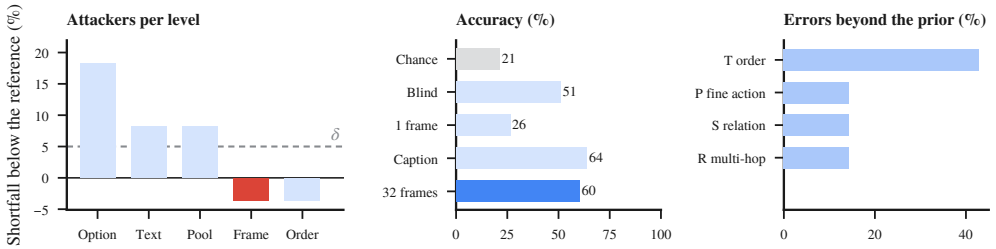
released 2025, chance 17%, reference 16% at 32 frames, breaks at the option level



Effective size 11 of 200, near-duplicates 100%, video dependence -18%, protocol issues 88% of errors, coverage pool 28, Video-Index 0

LongVideoBench

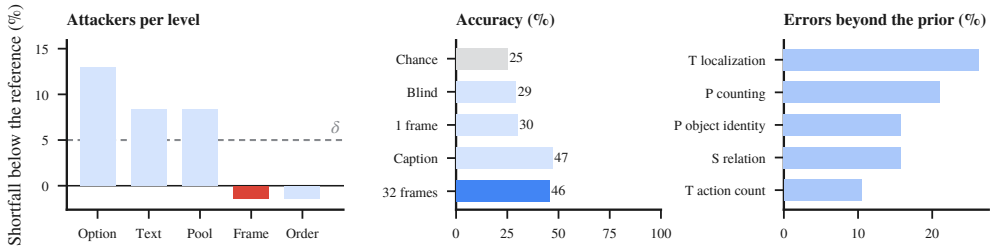
released 2024, chance 21%, reference 60% at 32 frames, breaks at the frame level



Effective size 55 of 200, near-duplicates 15%, video dependence +9%, protocol issues 59% of errors, coverage pool 221, Video-Index 2

LVBench

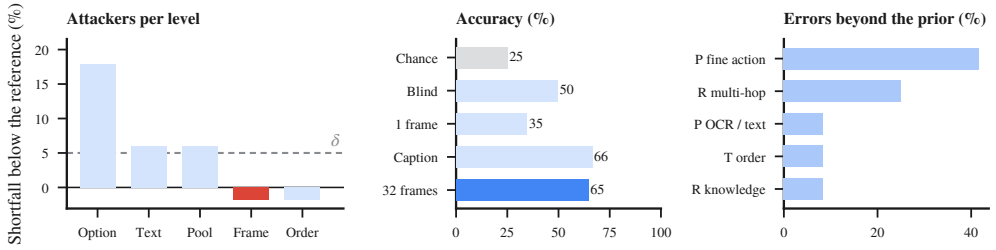
released 2024, chance 25%, reference 46% at 32 frames, breaks at the frame level



Effective size 30 of 200, near-duplicates 68%, video dependence +16%, protocol issues 40% of errors, coverage pool 70, Video-Index 8

MLVU

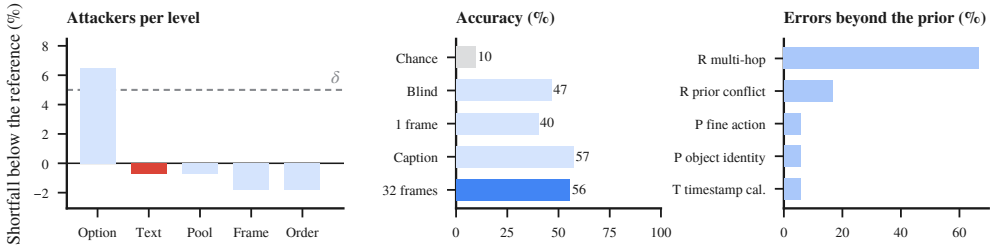
released 2024, chance 25%, reference 65% at 32 frames, breaks at the frame level



Effective size 22 of 200, near-duplicates 77%, video dependence +15%, protocol issues 59% of errors, coverage pool 111, Video-Index 6

MMR-V

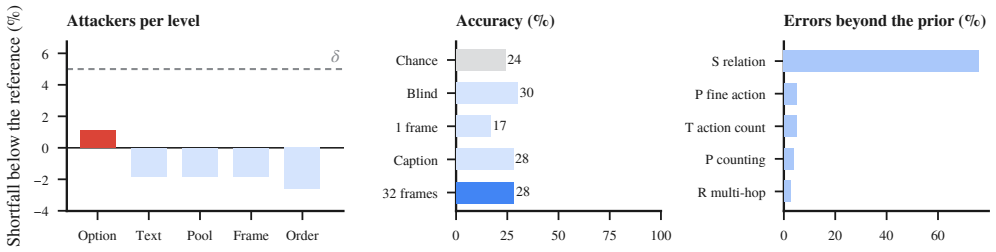
released 2025, chance 10%, reference 56% at 32 frames, breaks at the text level



Effective size 40 of 200, near-duplicates 24%, video dependence +9%, protocol issues 25% of errors, coverage pool 156, Video-Index 1

MMSI-Video-Bench

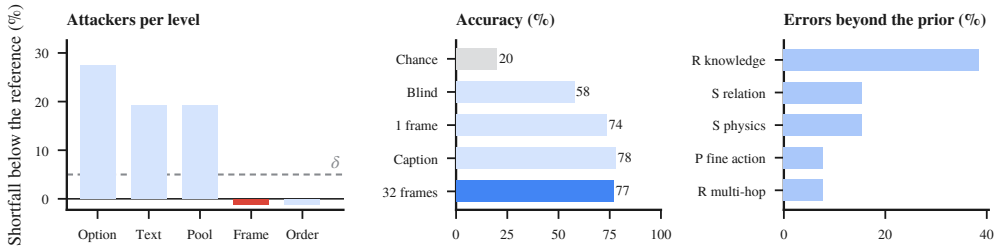
released 2025, chance 24%, reference 28% at 32 frames, breaks at the option level



Effective size 32 of 200, near-duplicates 48%, video dependence -2%, protocol issues 14% of errors, coverage pool 144, Video-Index 1

MMVU

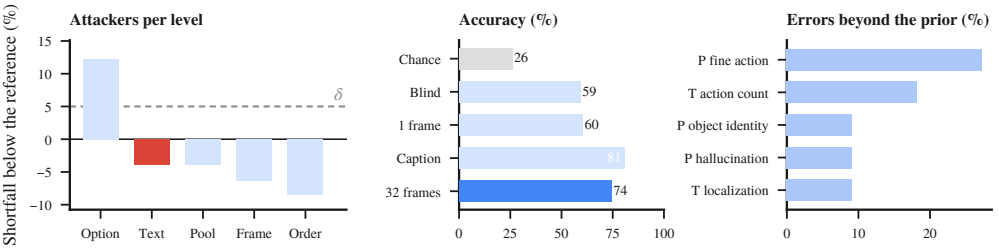
released 2025, chance 20%, reference 77% at 32 frames, breaks at the frame level



Effective size 50 of 200, near-duplicates 25%, video dependence +19%, protocol issues 27% of errors, coverage pool 89, Video-Index 1

MMWorld

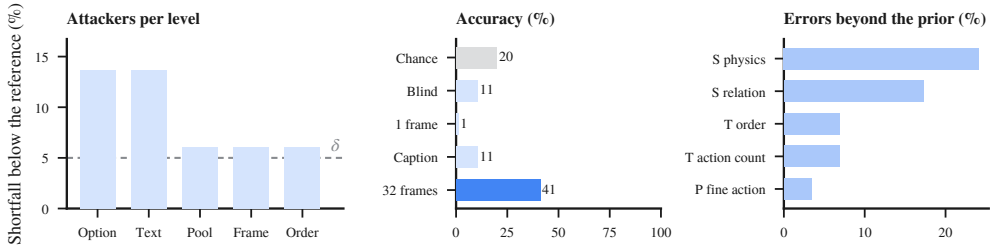
released 2024, chance 26%, reference 74% at 32 frames, breaks at the text level



Effective size 47 of 200, near-duplicates 53%, video dependence +15%, protocol issues 26% of errors, coverage pool 90, Video-Index 6

MORSE-500

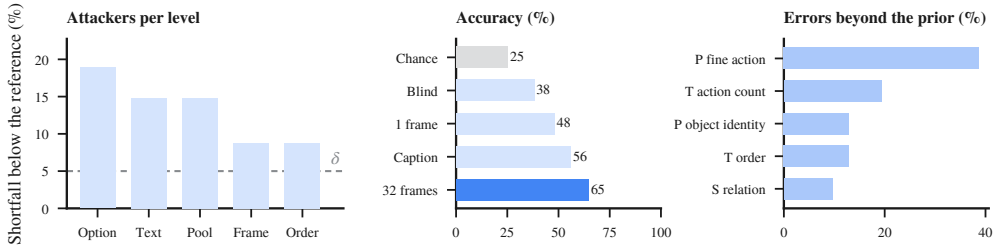
released 2025, chance 20%, reference 41% at 32 frames, survives every level



Effective size 5 of 200, near-duplicates 100%, video dependence +31%, protocol issues 16% of errors, coverage pool 0, Video-Index 0

MotionBench

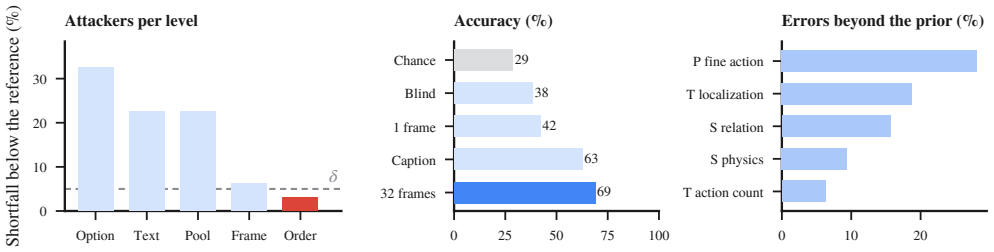
released 2025, chance 25%, reference 65% at 32 frames, survives every level



Effective size 44 of 200, near-duplicates 29%, video dependence +27%, protocol issues 9% of errors, coverage pool 219, Video-Index 11

MVBench

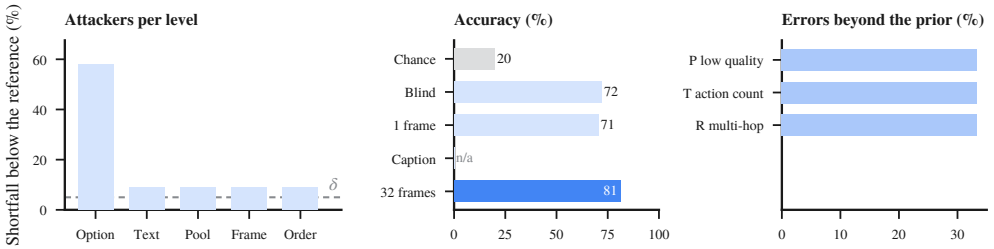
released 2023, chance 29%, reference 69% at 32 frames, breaks at the order level



Effective size 30 of 200, near-duplicates 67%, video dependence +30%, protocol issues 8% of errors, coverage pool 68, Video-Index 1

Neptune

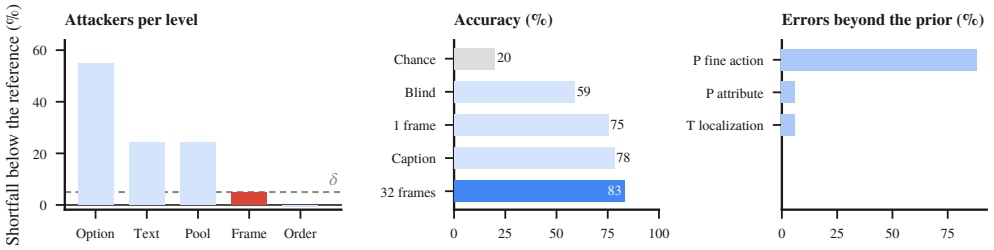
released 2024, chance 20%, reference 81% at 32 frames, survives every level



Effective size 56 of 200, near-duplicates 14%, video dependence +9%, protocol issues 73% of errors, coverage pool 0, Video-Index 0

NExT-QA

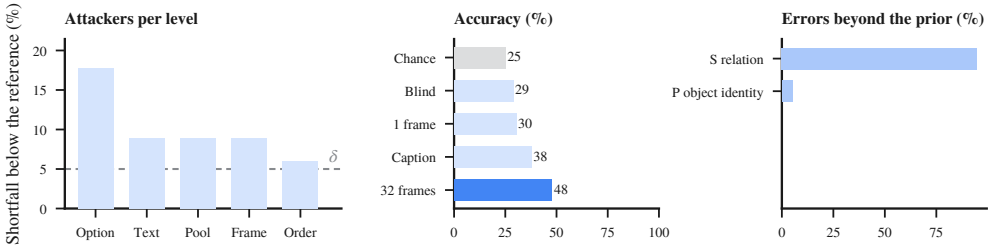
released 2021, chance 20%, reference 83% at 32 frames, breaks at the frame level



Effective size 52 of 200, near-duplicates 7%, video dependence +24%, protocol issues n/a of errors, coverage pool 84, Video-Index 2

OSI-Bench

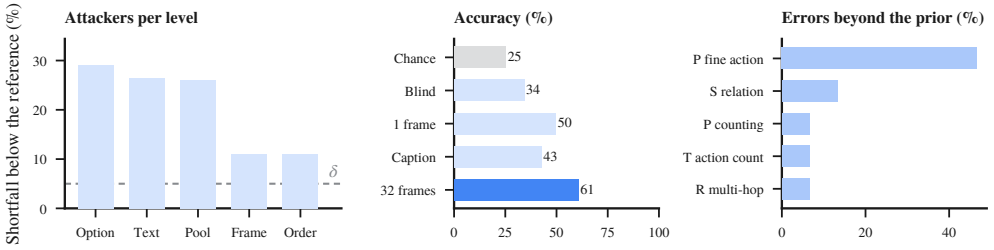
released 2026, chance 25%, reference 48% at 32 frames, survives every level



Effective size 16 of 200, near-duplicates 91%, video dependence +19%, protocol issues 4% of errors, coverage pool 22, Video-Index 1

PAI-Bench-U

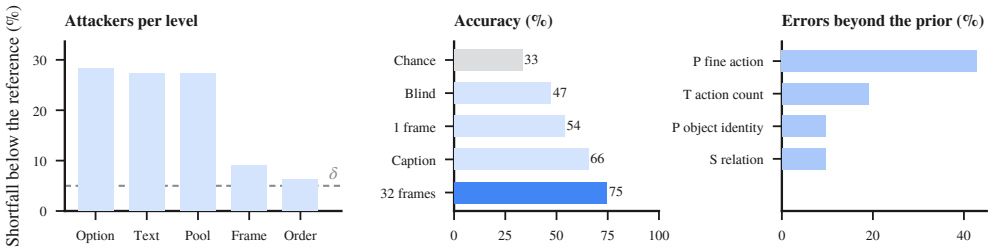
released 2026, chance 25%, reference 61% at 32 frames, survives every level



Effective size 44 of 200, near-duplicates 46%, video dependence +26%, protocol issues 0% of errors, coverage pool 1, Video-Index 0

Perception Test

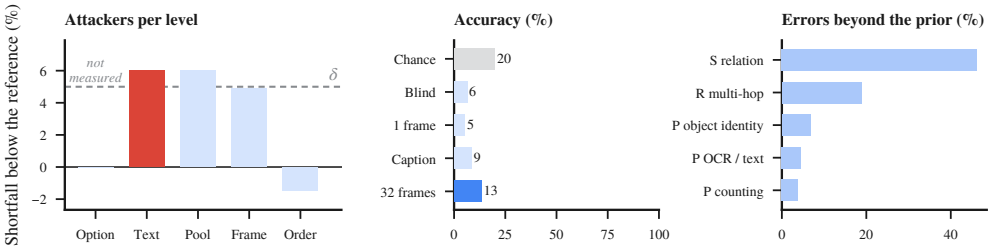
released 2023, chance 33%, reference 75% at 32 frames, survives every level



Effective size 32 of 200, near-duplicates 77%, video dependence +27%, protocol issues 4% of errors, coverage pool 101, Video-Index 19

PerceptionComp

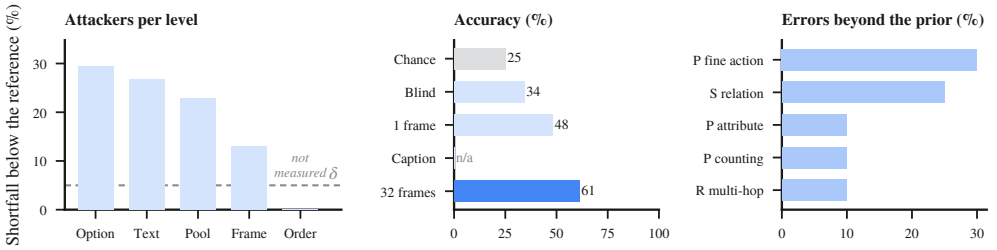
released 2026, chance 20%, reference 13% at 32 frames, breaks at the text level



Effective size 34 of 200, near-duplicates 67%, video dependence +7%, protocol issues 17% of errors, coverage pool 0, Video-Index 0

PhysicalAI

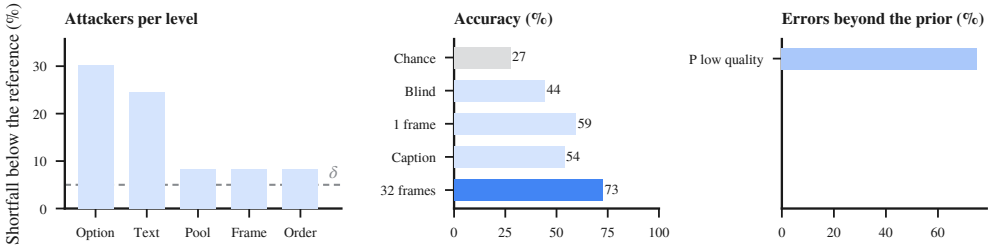
released 2026, chance 25%, reference 61% at 32 frames, survives every level



Effective size 44 of 200, near-duplicates 45%, video dependence +27%, protocol issues n/a of errors, coverage pool 0, Video-Index 0

Q-Bench-Video

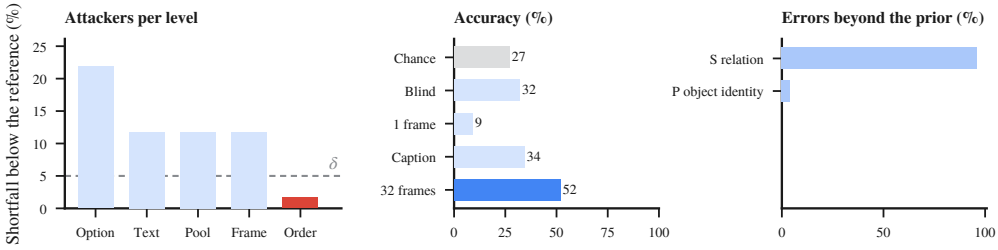
released 2024, chance 27%, reference 73% at 32 frames, survives every level



Effective size 38 of 200, near-duplicates 12%, video dependence +28%, protocol issues 38% of errors, coverage pool 69, Video-Index 0

ReVSI

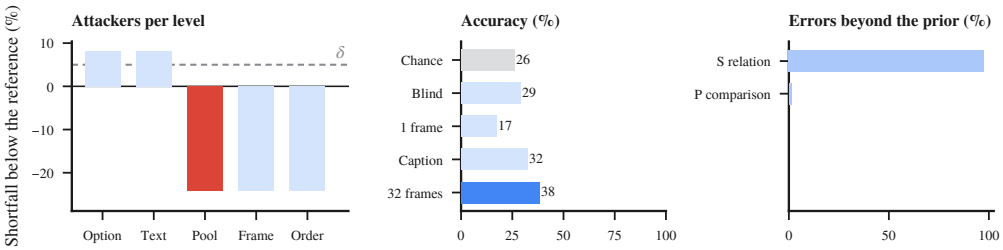
released 2026, chance 27%, reference 52% at 32 frames, breaks at the order level



Effective size 19 of 200, near-duplicates 94%, video dependence +20%, protocol issues 12% of errors, coverage pool 51, Video-Index 2

SAW-Bench

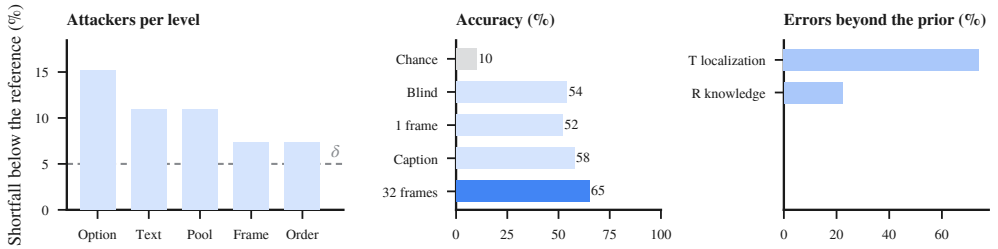
released 2026, chance 26%, reference 38% at 32 frames, breaks at the pool level



Effective size 8 of 200, near-duplicates 97%, video dependence +9%, protocol issues 0% of errors, coverage pool 28, Video-Index 3

SciVideoBench

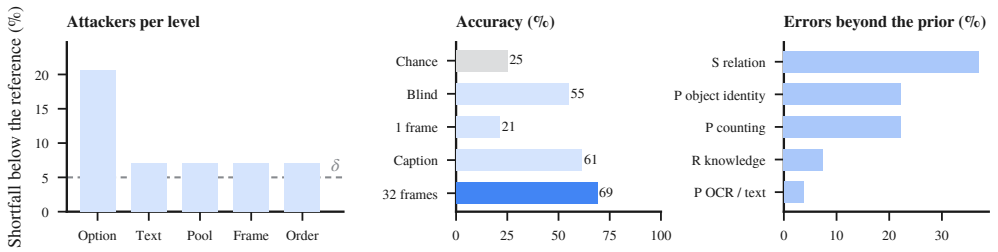
released 2025, chance 10%, reference 65% at 32 frames, survives every level



Effective size 16 of 200, near-duplicates 76%, video dependence +11%, protocol issues 2% of errors, coverage pool 33, Video-Index 0

SFI-Bench

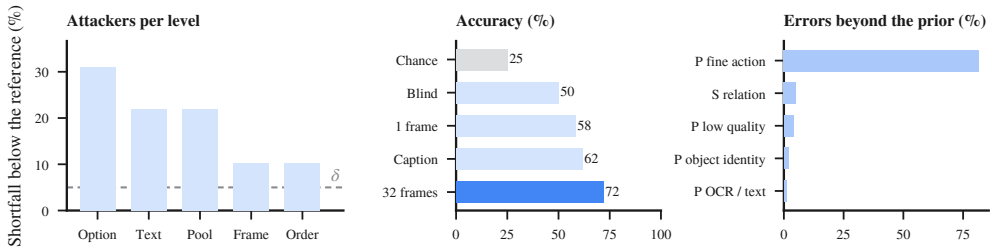
released 2026, chance 25%, reference 69% at 32 frames, survives every level



Effective size 28 of 200, near-duplicates 59%, video dependence +14%, protocol issues 3% of errors, coverage pool 20, Video-Index 3

SIS-Bench

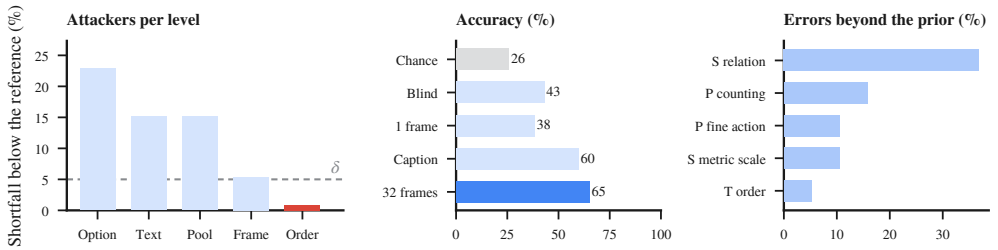
released 2026, chance 25%, reference 72% at 32 frames, survives every level



Effective size 11 of 200, near-duplicates 85%, video dependence +22%, protocol issues 0% of errors, coverage pool 89, Video-Index 5

SITE-Bench

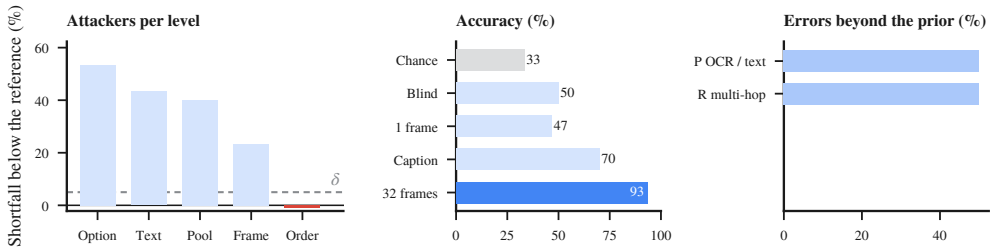
released 2025, chance 26%, reference 65% at 32 frames, breaks at the order level



Effective size 41 of 200, near-duplicates 47%, video dependence +22%, protocol issues 9% of errors, coverage pool 184, Video-Index 2

SoMi-ToM

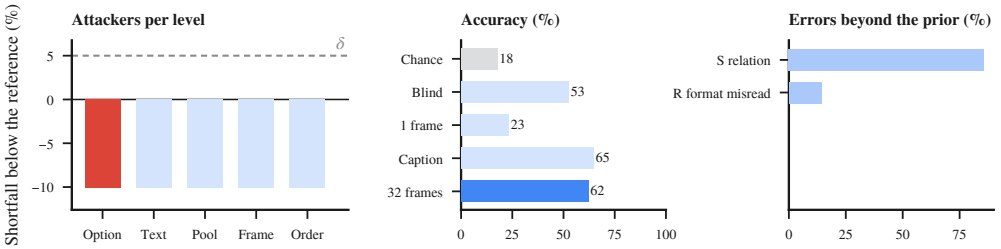
released 2025, chance 33%, reference 93% at 32 frames, breaks at the order level



Effective size 3 of 200, near-duplicates 100%, video dependence +43%, protocol issues 0% of errors, coverage pool 0, Video-Index 0

SpatialTree

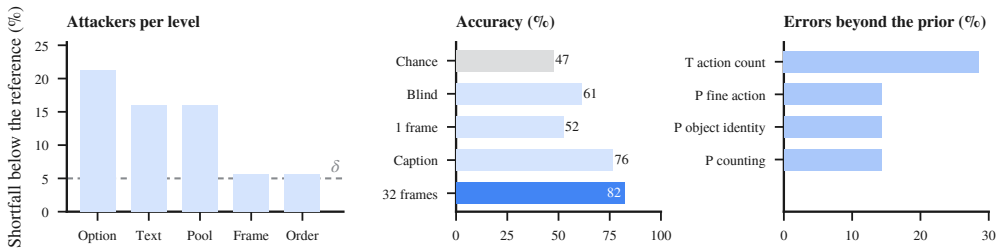
released 2025, chance 18%, reference 62% at 32 frames, breaks at the option level



Effective size 16 of 200, near-duplicates 79%, video dependence +10%, protocol issues 16% of errors, coverage pool 9, Video-Index 0

ST-Evidence

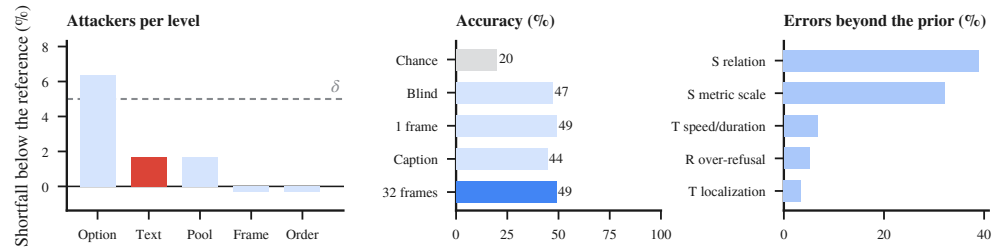
released 2026, chance 47%, reference 82% at 32 frames, survives every level



Effective size 48 of 200, near-duplicates 35%, video dependence +21%, protocol issues 15% of errors, coverage pool 62, Video-Index 3

STI-Bench

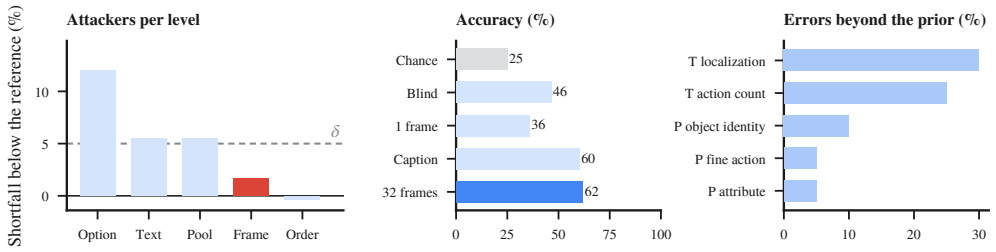
released 2025, chance 20%, reference 49% at 32 frames, breaks at the text level



Effective size 19 of 200, near-duplicates 88%, video dependence +2%, protocol issues 1% of errors, coverage pool 128, Video-Index 0

StreamingBench

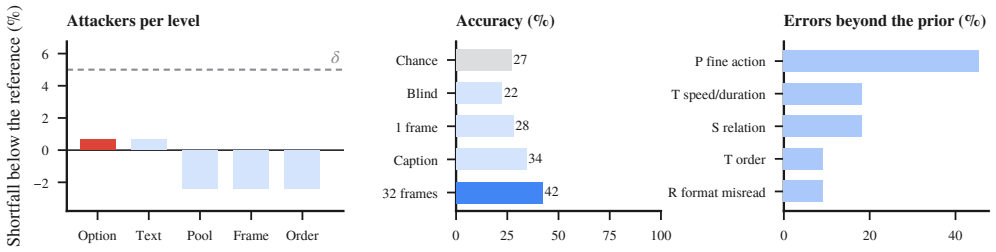
released 2024, chance 25%, reference 62% at 32 frames, breaks at the frame level



Effective size 41 of 200, near-duplicates 54%, video dependence +15%, protocol issues 31% of errors, coverage pool 203, Video-Index 9

TempCompass

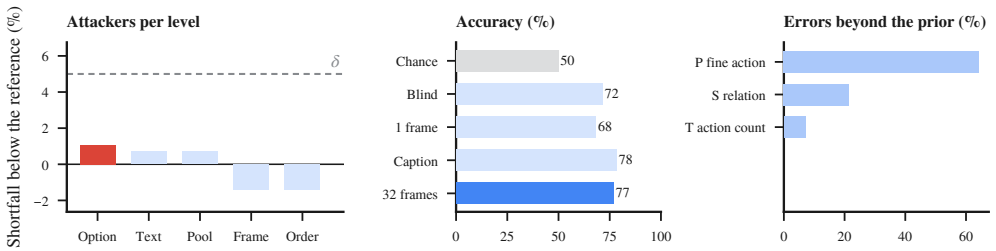
released 2024, chance 27%, reference 42% at 32 frames, breaks at the option level



Effective size 34 of 200, near-duplicates 58%, video dependence +20%, protocol issues 57% of errors, coverage pool 75, Video-Index 13

TemporalBench

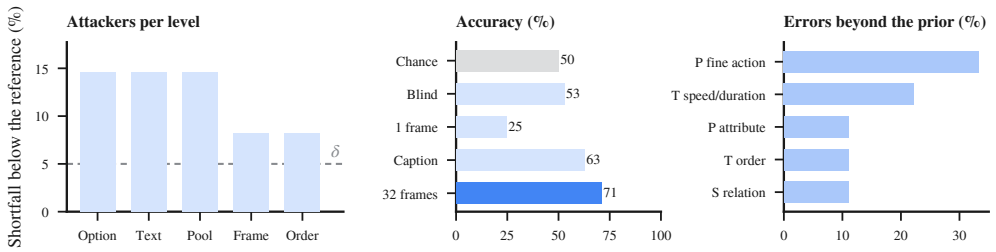
released 2024, chance 50%, reference 77% at 32 frames, breaks at the option level



Effective size 29 of 200, near-duplicates 62%, video dependence +5%, protocol issues 5% of errors, coverage pool 179, Video-Index 12

TimeBlind

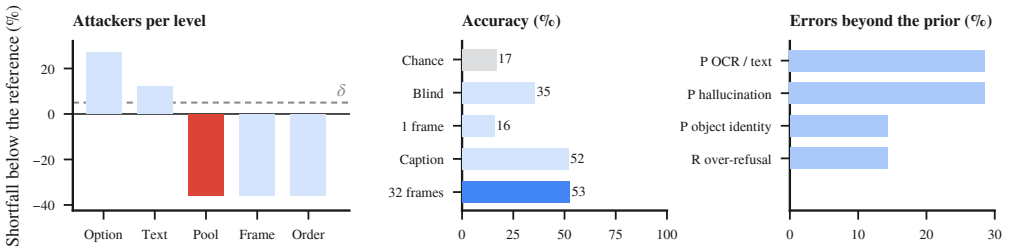
released 2026, chance 50%, reference 71% at 32 frames, survives every level



Effective size 33 of 200, near-duplicates 80%, video dependence +18%, protocol issues 6% of errors, coverage pool 33, Video-Index 9

TimeScope

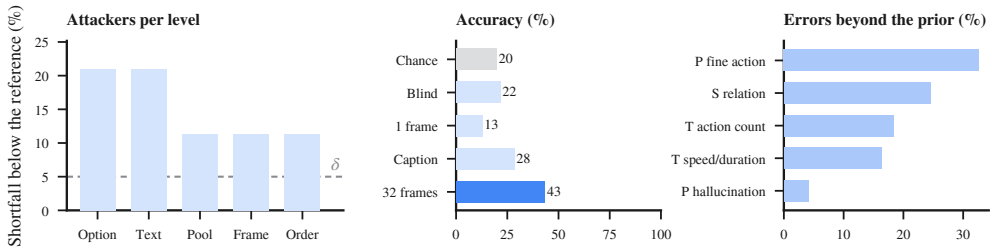
released 2025, chance 17%, reference 53% at 32 frames, breaks at the pool level



Effective size 13 of 200, near-duplicates 100%, video dependence +17%, protocol issues 82% of errors, coverage pool 96, Video-Index 0

TOMATO

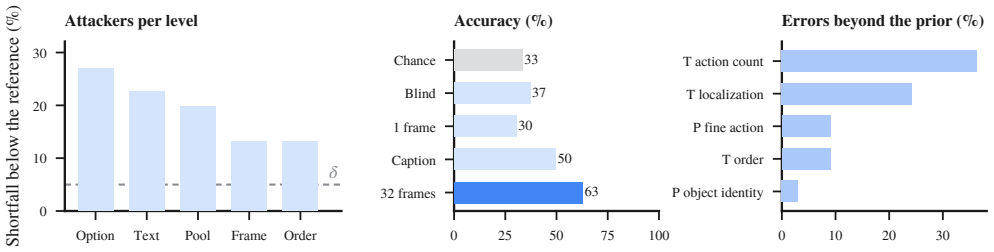
released 2024, chance 20%, reference 43% at 32 frames, survives every level



Effective size 29 of 200, near-duplicates 72%, video dependence +22%, protocol issues 10% of errors, coverage pool 142, Video-Index 4

TVBench

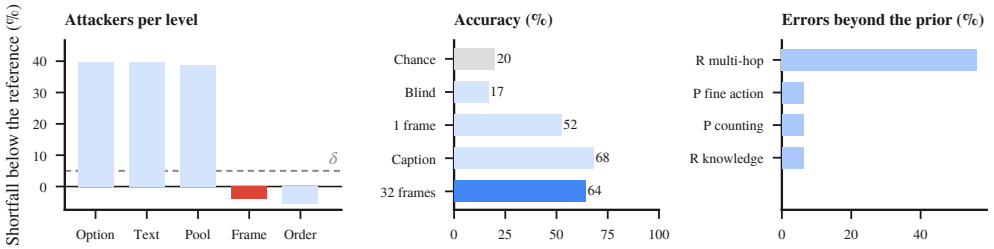
released 2024, chance 33%, reference 63% at 32 frames, survives every level



Effective size 18 of 200, near-duplicates 90%, video dependence +26%, protocol issues 13% of errors, coverage pool 35, Video-Index 3

v-HUB

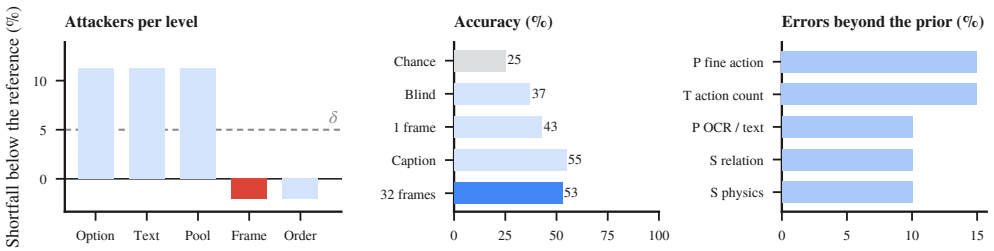
released 2025, chance 20%, reference 64% at 32 frames, breaks at the frame level



Effective size 42 of 200, near-duplicates 22%, video dependence +48%, protocol issues 50% of errors, coverage pool 74, Video-Index 0

V2P-Bench

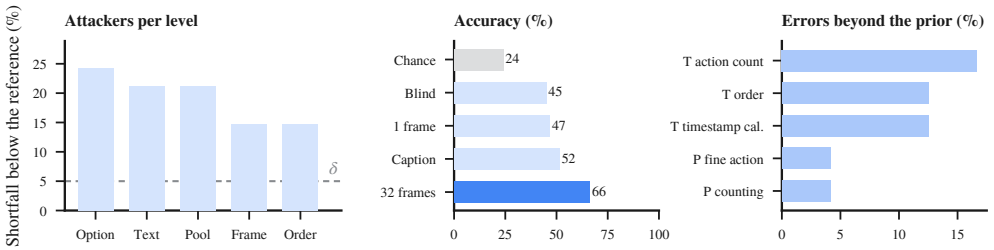
released 2025, chance 25%, reference 53% at 32 frames, breaks at the frame level



Effective size 39 of 200, near-duplicates 25%, video dependence +16%, protocol issues 49% of errors, coverage pool 161, Video-Index 2

VCR-Bench

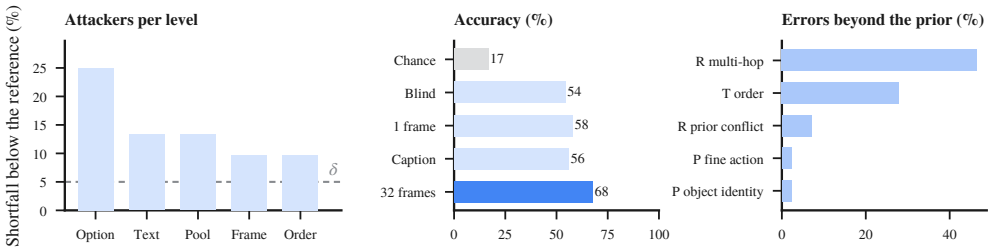
released 2025, chance 24%, reference 66% at 32 frames, survives every level



Effective size 53 of 200, near-duplicates 9%, video dependence +21%, protocol issues 2% of errors, coverage pool 81, Video-Index 1

Video-Holmes

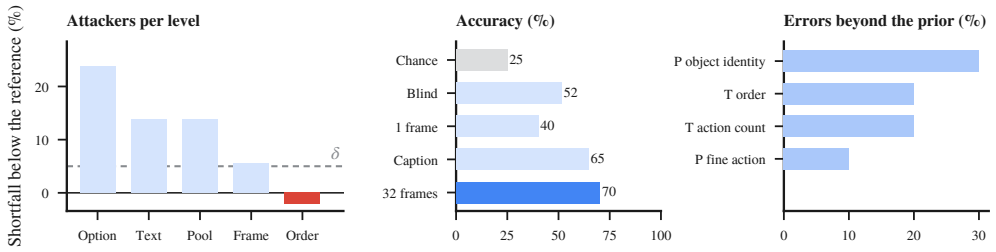
released 2025, chance 17%, reference 68% at 32 frames, survives every level



Effective size 32 of 200, near-duplicates 36%, video dependence +13%, protocol issues 4% of errors, coverage pool 69, Video-Index 21

Video-MME

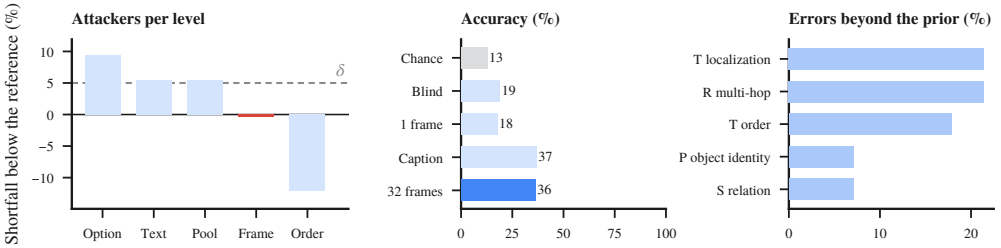
released 2024, chance 25%, reference 70% at 32 frames, breaks at the order level



Effective size 53 of 200, near-duplicates 5%, video dependence +19%, protocol issues 40% of errors, coverage pool 308, Video-Index 9

Video-MME-v2

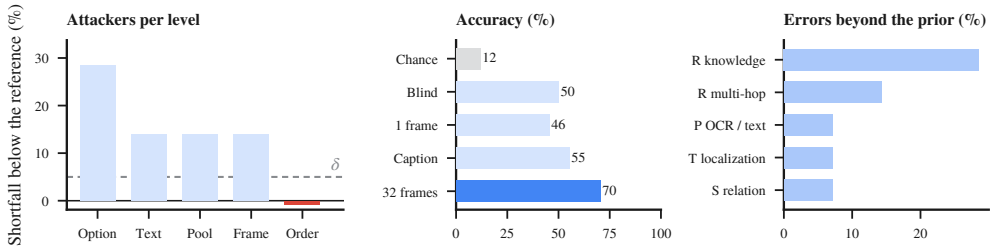
released 2026, chance 13%, reference 36% at 32 frames, breaks at the frame level



Effective size 62 of 200, near-duplicates 15%, video dependence +17%, protocol issues 28% of errors, coverage pool 501, Video-Index 28

Video-MMMU

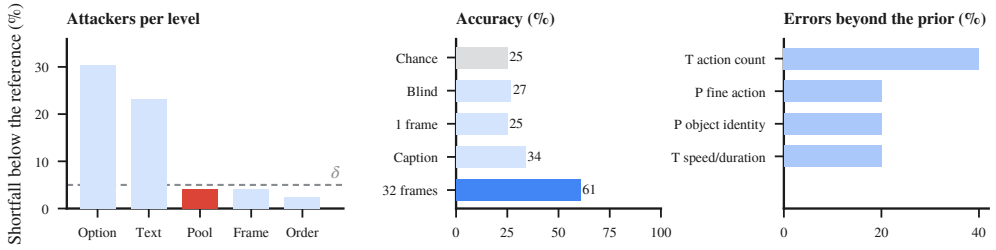
released 2025, chance 12%, reference 70% at 32 frames, breaks at the order level



Effective size 60 of 200, near-duplicates 23%, video dependence +20%, protocol issues 38% of errors, coverage pool 95, Video-Index 1

Video-TT

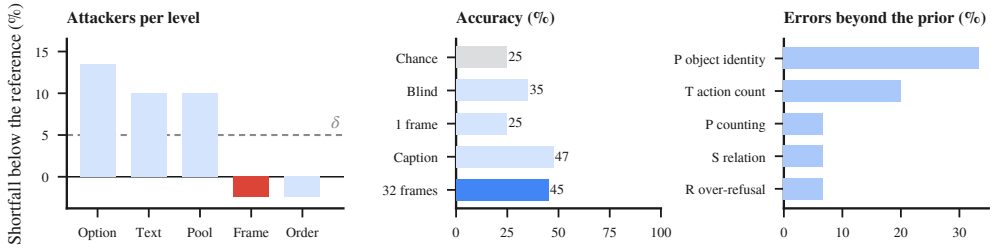
released 2025, chance 25%, reference 61% at 32 frames, breaks at the pool level



Effective size 62 of 200, near-duplicates 13%, video dependence +34%, protocol issues 14% of errors, coverage pool 172, Video-Index 9

VideoEval-Pro

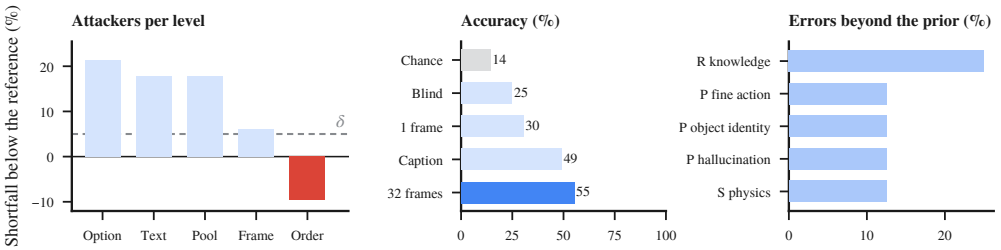
released 2025, chance 25%, reference 45% at 32 frames, breaks at the frame level



Effective size 59 of 200, near-duplicates 17%, video dependence +10%, protocol issues 47% of errors, coverage pool 51, Video-Index 3

VideoKR

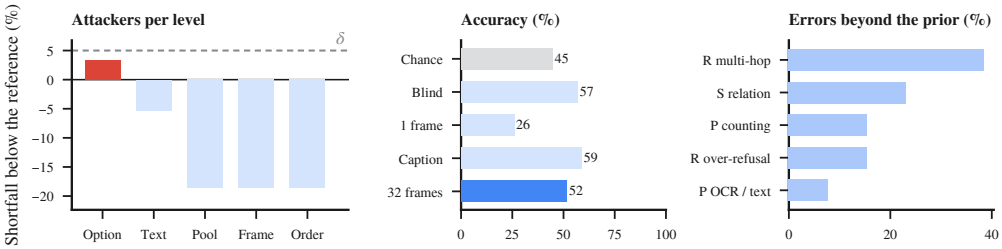
released 2026, chance 14%, reference 55% at 32 frames, breaks at the order level



Effective size 50 of 200, near-duplicates 26%, video dependence +30%, protocol issues 52% of errors, coverage pool 59, Video-Index 0

VideoMathQA

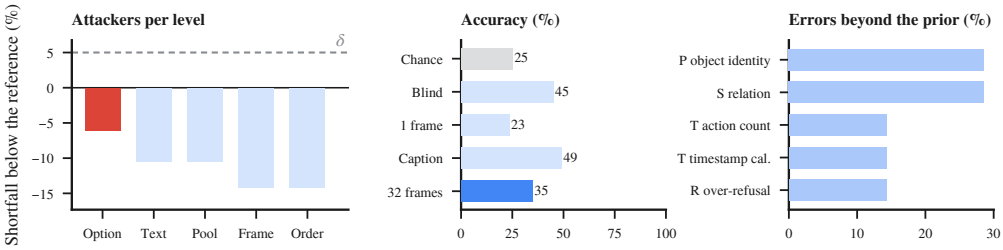
released 2025, chance 45%, reference 52% at 32 frames, breaks at the option level



Effective size 40 of 200, near-duplicates 63%, video dependence -5%, protocol issues 43% of errors, coverage pool 76, Video-Index 2

VideoOdyssey

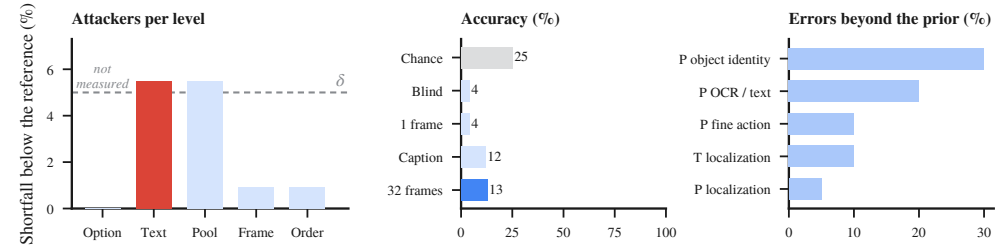
released 2026, chance 25%, reference 35% at 32 frames, breaks at the option level



Effective size 42 of 200, near-duplicates 27%, video dependence -10%, protocol issues 51% of errors, coverage pool 105, Video-Index 26

VideoSIAH-Eval

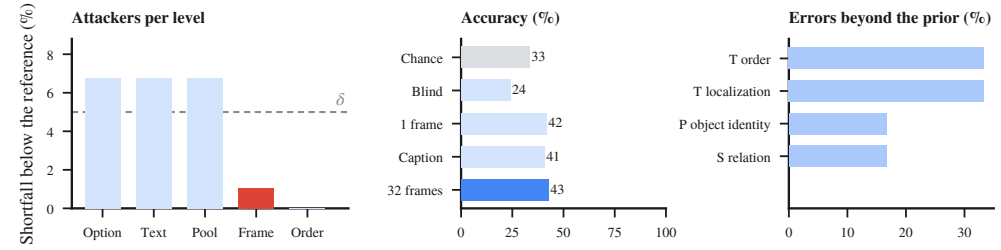
released 2025, chance 25%, reference 13% at 32 frames, breaks at the text level



Effective size 64 of 200, near-duplicates 8%, video dependence +9%, protocol issues 82% of errors, coverage pool 0, Video-Index 0

VidPair-Halluc

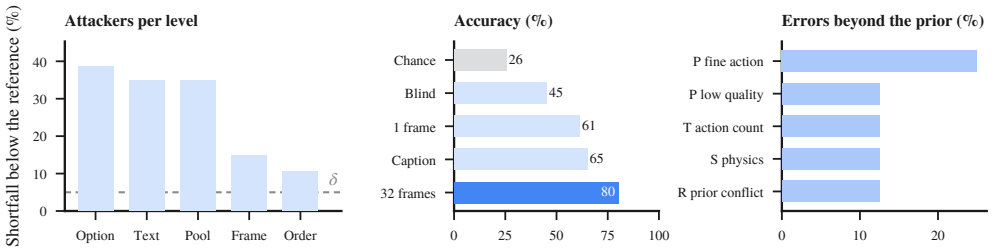
released 2026, chance 33%, reference 43% at 32 frames, breaks at the frame level



Effective size 27 of 200, near-duplicates 75%, video dependence +19%, protocol issues 31% of errors, coverage pool 37, Video-Index 2

ViMUL-Bench

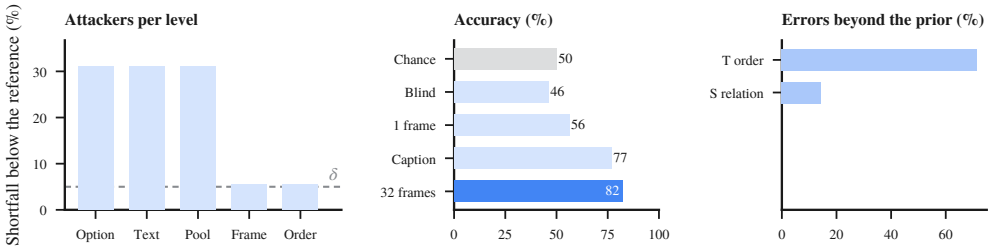
released 2025, chance 26%, reference 80% at 32 frames, survives every level



Effective size 57 of 200, near-duplicates 10%, video dependence +35%, protocol issues 49% of errors, coverage pool 244, Video-Index 7

Vinoground

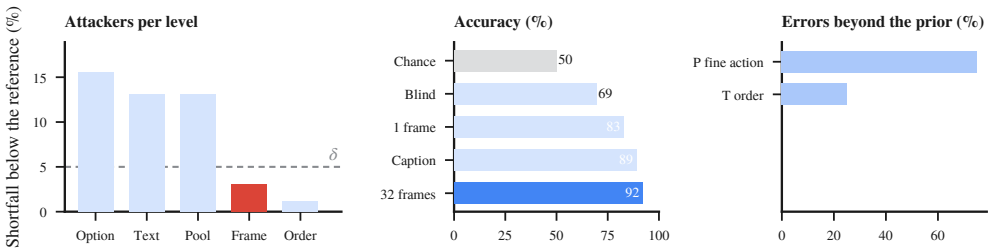
released 2024, chance 50%, reference 82% at 32 frames, survives every level



Effective size 12 of 200, near-duplicates 81%, video dependence +36%, protocol issues 0% of errors, coverage pool 5, Video-Index 18

VITATECS

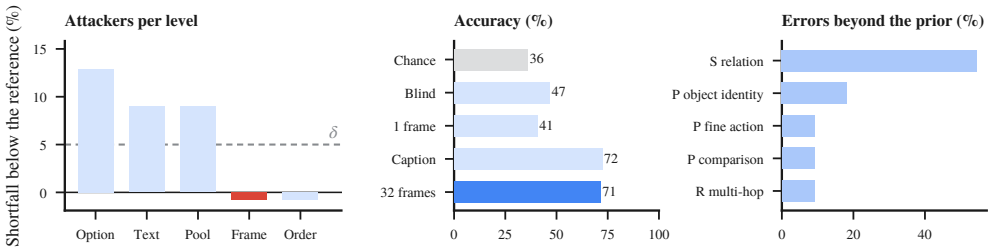
released 2023, chance 50%, reference 92% at 32 frames, breaks at the frame level



Effective size 12 of 200, near-duplicates 79%, video dependence +23%, protocol issues 14% of errors, coverage pool 7, Video-Index 24

VKnowU

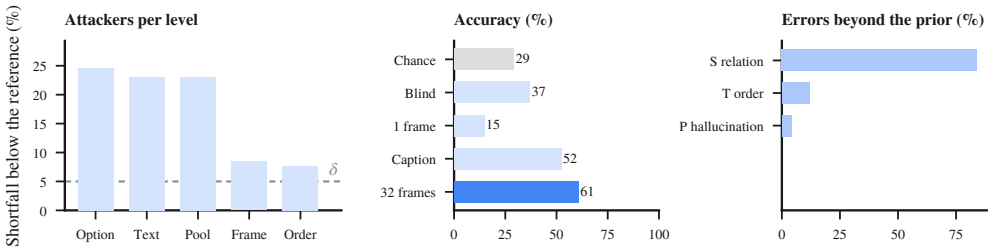
released 2025, chance 36%, reference 71% at 32 frames, breaks at the frame level



Effective size 45 of 200, near-duplicates 42%, video dependence +25%, protocol issues 29% of errors, coverage pool 124, Video-Index 6

VSI-Bench

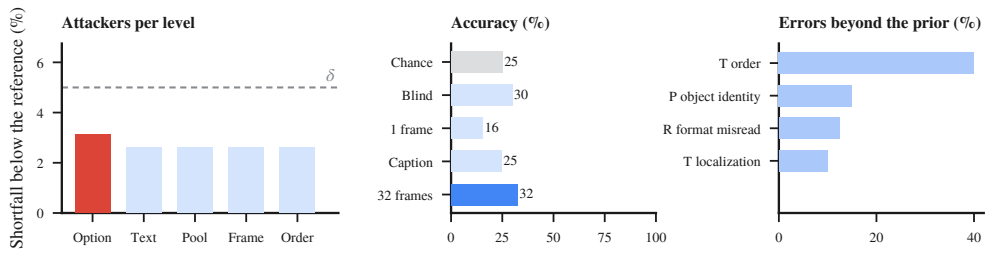
released 2024, chance 29%, reference 61% at 32 frames, survives every level



Effective size 15 of 200, near-duplicates 98%, video dependence +24%, protocol issues 8% of errors, coverage pool 21, Video-Index 0

VSI-Super-Wild

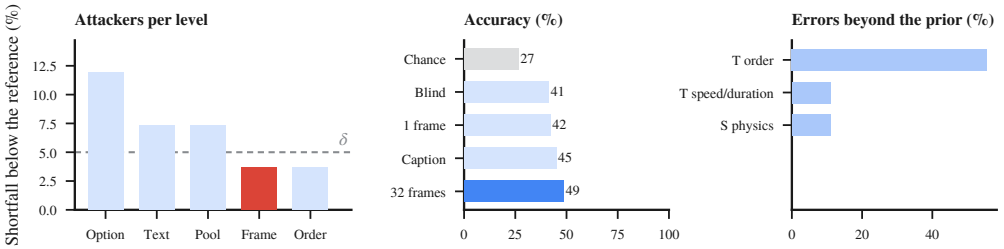
released 2026, chance 25%, reference 32% at 32 frames, breaks at the option level



Effective size 9 of 200, near-duplicates 97%, video dependence +3%, protocol issues 15% of errors, coverage pool 116, Video-Index 0

VSTAT

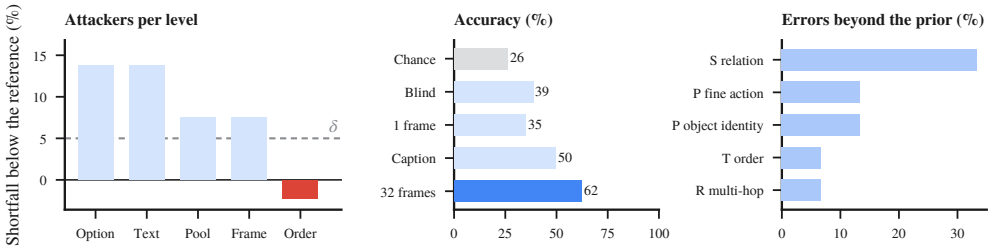
released 2026, chance 27%, reference 49% at 32 frames, breaks at the frame level



Effective size 12 of 200, near-duplicates 100%, video dependence +7%, protocol issues n/a of errors, coverage pool 32, Video-Index 0

VURB

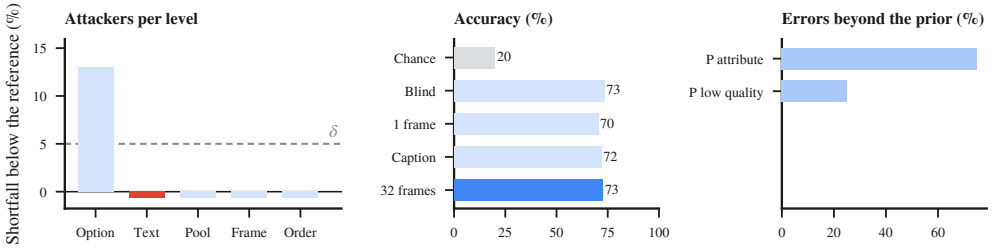
released 2026, chance 26%, reference 62% at 32 frames, breaks at the order level



Effective size 44 of 200, near-duplicates 61%, video dependence +24%, protocol issues 42% of errors, coverage pool 17, Video-Index 0

ALLVB

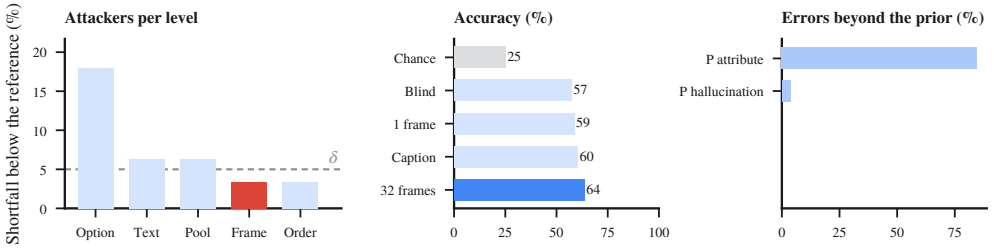
released 2025, chance 20%, reference 73% at 32 frames, breaks at the text level



Effective size 34 of 200, near-duplicates 24%, video dependence -1%, protocol issues 9% of errors, coverage pool 139, Video-Index 45

Causal2Needles

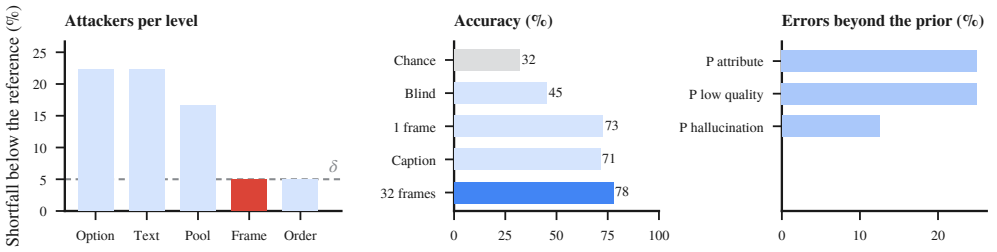
released 2025, chance 25%, reference 64% at 32 frames, breaks at the frame level



Effective size 27 of 200, near-duplicates 42%, video dependence +6%, protocol issues 0% of errors, coverage pool 26, Video-Index 0

CAViAR

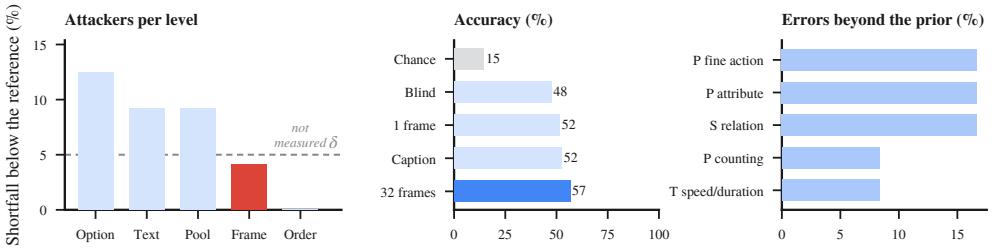
released 2026, chance 32%, reference 78% at 32 frames, breaks at the frame level



Effective size 4 of 200, near-duplicates 100%, video dependence +32%, protocol issues 38% of errors, coverage pool 10, Video-Index 0

CG-Bench

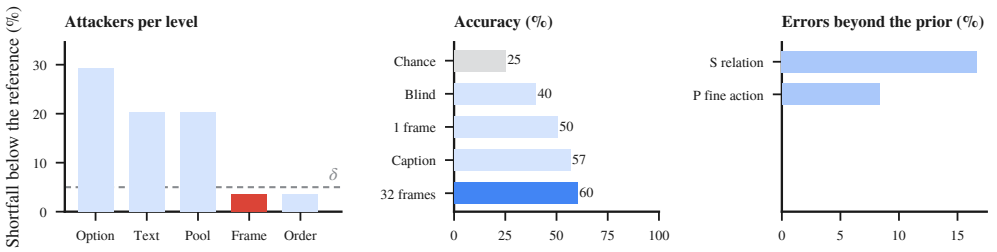
released 2024, chance 15%, reference 57% at 32 frames, breaks at the frame level



Effective size n/a of 200, near-duplicates n/a, video dependence +9%, protocol issues 8% of errors, coverage pool 0, Video-Index 0

CineTechBench

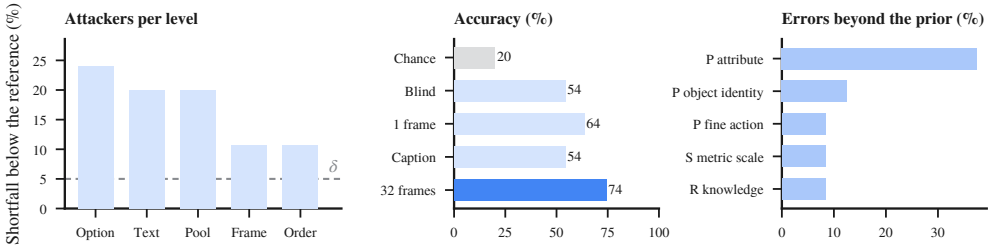
released 2025, chance 25%, reference 60% at 32 frames, breaks at the frame level



Effective size 8 of 200, near-duplicates 88%, video dependence +20%, protocol issues 0% of errors, coverage pool 5, Video-Index 0

Colon-Bench

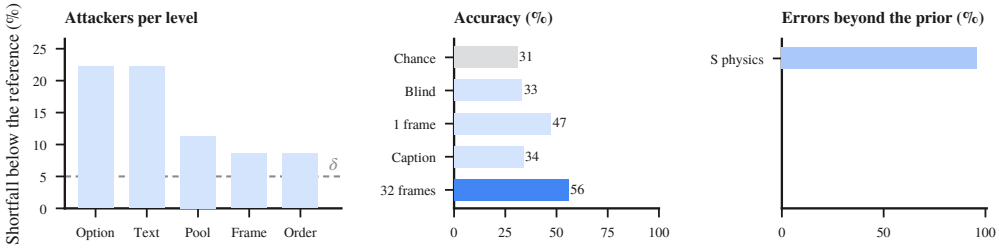
released 2026, chance 20%, reference 74% at 32 frames, survives every level



Effective size 8 of 200, near-duplicates 100%, video dependence +20%, protocol issues 0% of errors, coverage pool 45, Video-Index 1

ContPhy

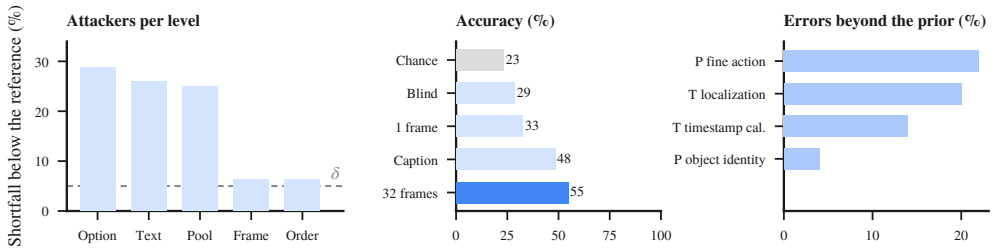
released 2024, chance 31%, reference 56% at 32 frames, survives every level



Effective size 9 of 200, near-duplicates 100%, video dependence +23%, protocol issues 0% of errors, coverage pool 11, Video-Index 0

E.T. Bench

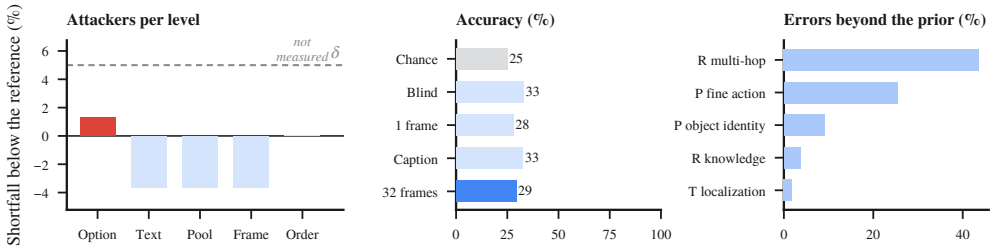
released 2024, chance 23%, reference 55% at 32 frames, survives every level



Effective size 12 of 200, near-duplicates 97%, video dependence +26%, protocol issues 0% of errors, coverage pool 26, Video-Index 2

EgoPlan-Bench

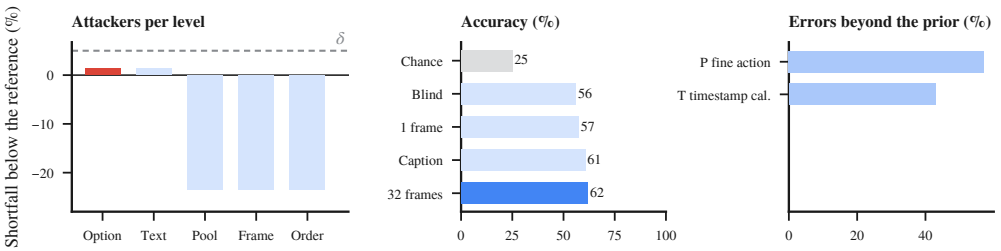
released 2023, chance 25%, reference 29% at 32 frames, breaks at the option level



Effective size 9 of 200, near-duplicates 97%, video dependence -4%, protocol issues 12% of errors, coverage pool 0, Video-Index 0

FineBench

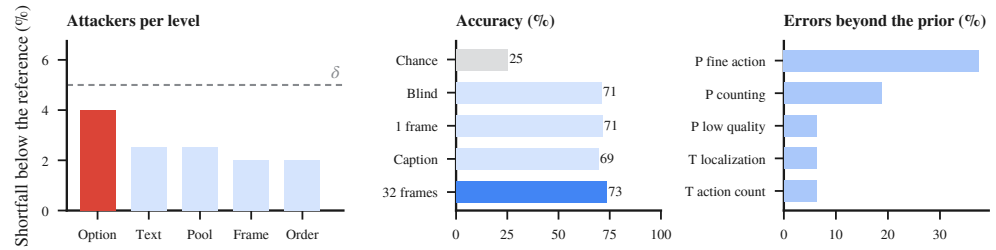
released 2026, chance 25%, reference 62% at 32 frames, breaks at the option level



Effective size 6 of 200, near-duplicates 100%, video dependence +6%, protocol issues 0% of errors, coverage pool 4, Video-Index 0

ForeSeaQA

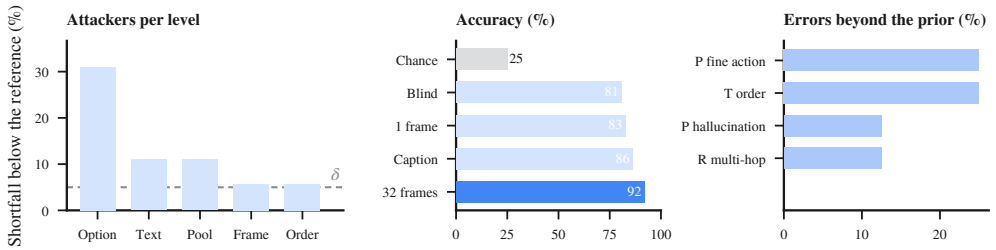
released 2026, chance 25%, reference 73% at 32 frames, breaks at the option level



Effective size 27 of 200, near-duplicates 84%, video dependence +3%, protocol issues 0% of errors, coverage pool 56, Video-Index 1

GUI-World

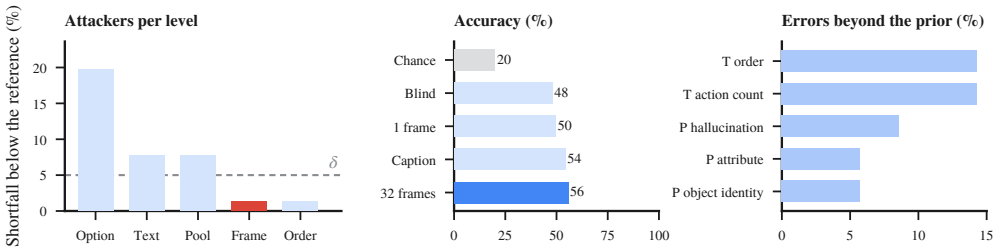
released 2025, chance 25%, reference 92% at 32 frames, survives every level



Effective size 54 of 200, near-duplicates 12%, video dependence +11%, protocol issues 5% of errors, coverage pool 79, Video-Index 6

HERBench

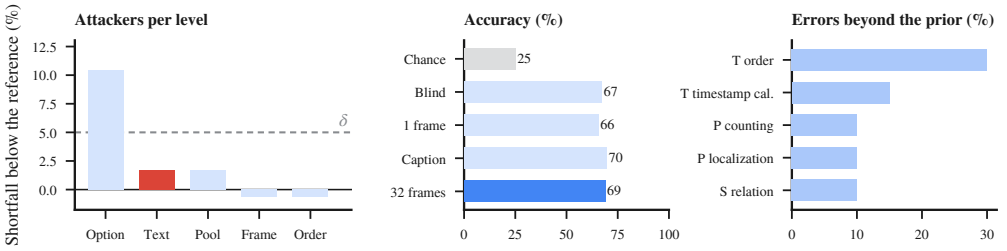
released 2026, chance 20%, reference 56% at 32 frames, breaks at the frame level



Effective size 12 of 200, near-duplicates 99%, video dependence +8%, protocol issues 0% of errors, coverage pool 48, Video-Index 0

HLV-1K

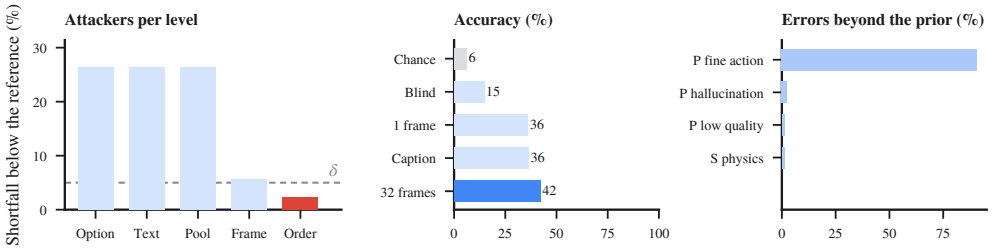
released 2025, chance 25%, reference 69% at 32 frames, breaks at the text level



Effective size 33 of 200, near-duplicates 46%, video dependence +2%, protocol issues 4% of errors, coverage pool 421, Video-Index 7

iSafetyBench

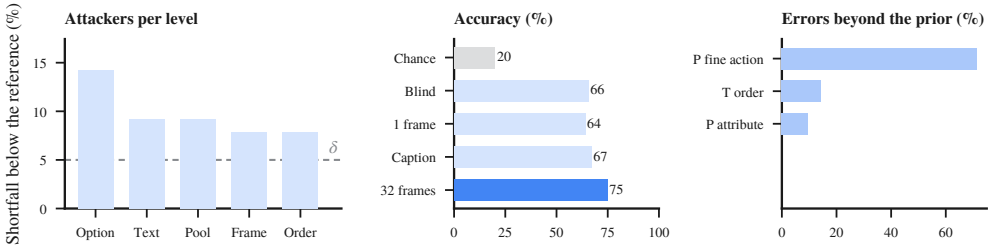
released 2025, chance 6%, reference 42% at 32 frames, breaks at the order level



Effective size 10 of 200, near-duplicates 99%, video dependence +27%, protocol issues 4% of errors, coverage pool 13, Video-Index 0

K9-Bench

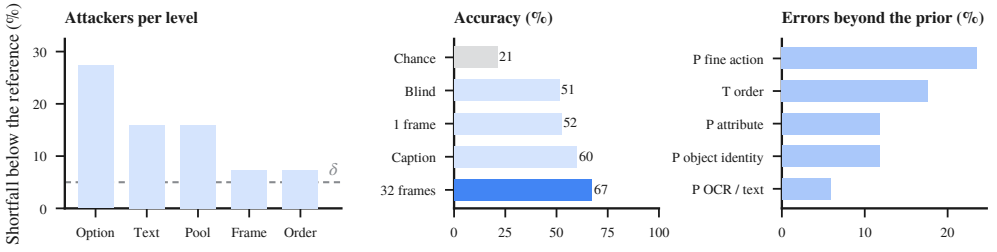
released 2026, chance 20%, reference 75% at 32 frames, survives every level



Effective size 29 of 200, near-duplicates 41%, video dependence +9%, protocol issues 0% of errors, coverage pool 305, Video-Index 21

KFS-Bench

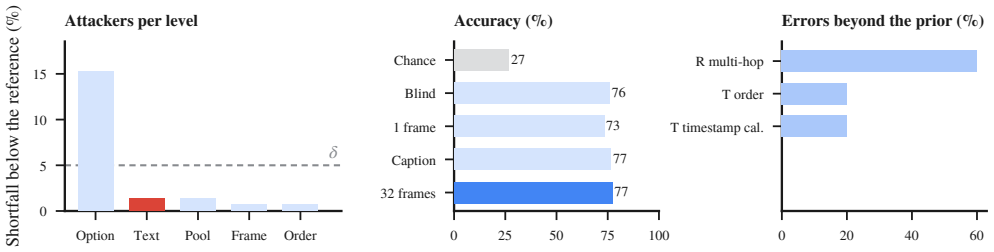
released 2026, chance 21%, reference 67% at 32 frames, survives every level



Effective size 51 of 200, near-duplicates 27%, video dependence +16%, protocol issues 7% of errors, coverage pool 13, Video-Index 0

LongVideo-Reason

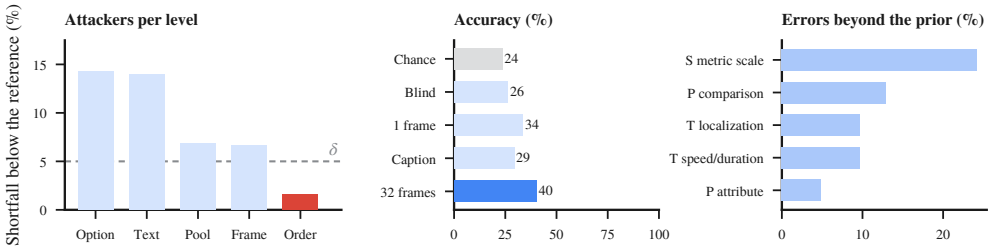
released 2025, chance 27%, reference 77% at 32 frames, breaks at the text level



Effective size 43 of 200, near-duplicates 11%, video dependence +1%, protocol issues 8% of errors, coverage pool 87, Video-Index 1

M3-Verse

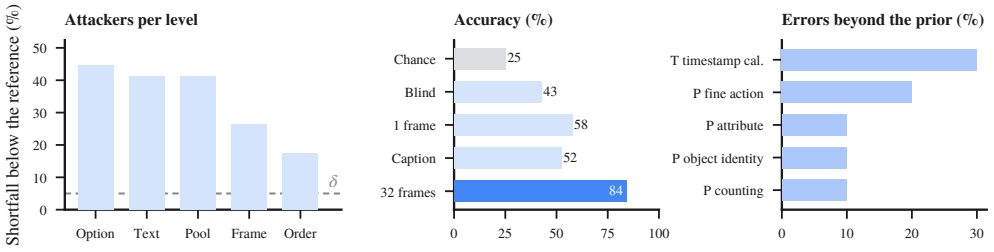
released 2025, chance 24%, reference 40% at 32 frames, breaks at the order level



Effective size 4 of 200, near-duplicates 100%, video dependence +14%, protocol issues 1% of errors, coverage pool 24, Video-Index 1

MME-VideoOCR

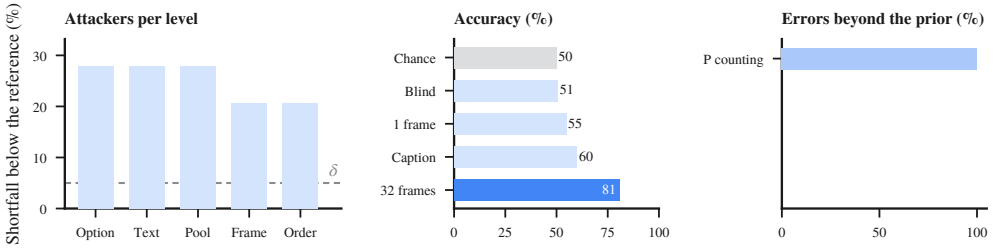
released 2025, chance 25%, reference 84% at 32 frames, survives every level



Effective size 70 of 200, near-duplicates 10%, video dependence +41%, protocol issues 3% of errors, coverage pool 184, Video-Index 5

NaturalBench-Video

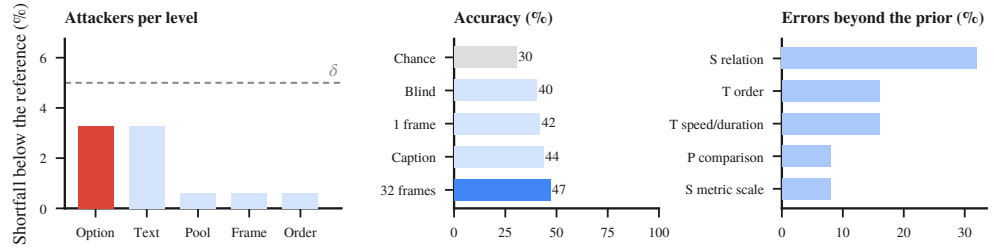
released 2025, chance 50%, reference 81% at 32 frames, survives every level



Effective size 44 of 200, near-duplicates 67%, video dependence +30%, protocol issues 0% of errors, coverage pool 38, Video-Index 7

OVO-S-Bench

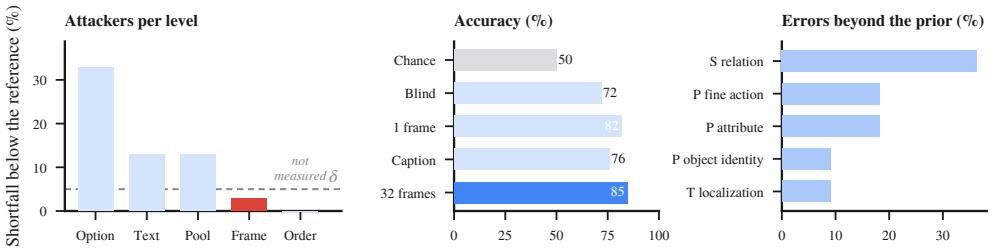
released 2026, chance 30%, reference 47% at 32 frames, breaks at the option level



Effective size 16 of 200, near-duplicates 91%, video dependence +7%, protocol issues 2% of errors, coverage pool 125, Video-Index 2

PLM-VideoBench

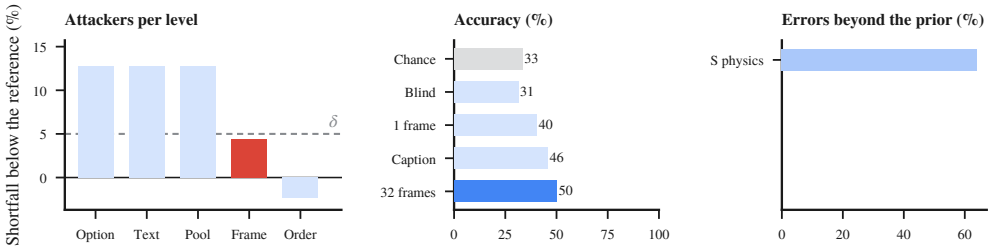
released 2025, chance 50%, reference 85% at 32 frames, breaks at the frame level



Effective size 48 of 200, near-duplicates 19%, video dependence +13%, protocol issues 0% of errors, coverage pool 0, Video-Index 0

ReactHuman

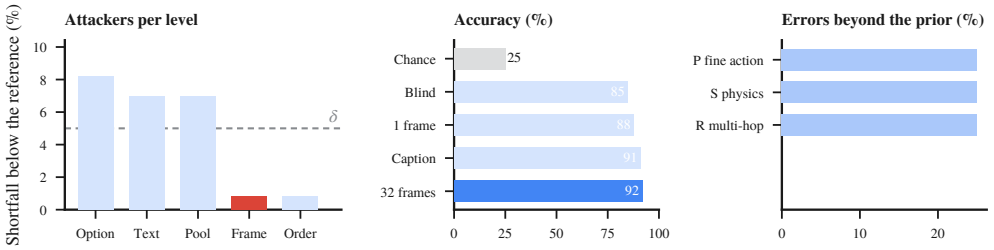
released 2026, chance 33%, reference 50% at 32 frames, breaks at the frame level



Effective size 2 of 200, near-duplicates 100%, video dependence +19%, protocol issues 0% of errors, coverage pool 17, Video-Index 0

ReXTime

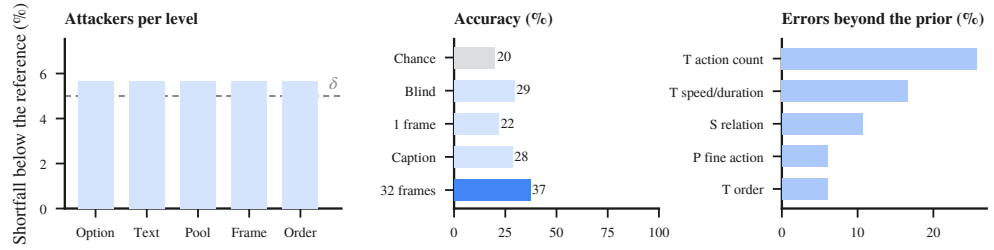
released 2024, chance 25%, reference 92% at 32 frames, breaks at the frame level



Effective size 62 of 200, near-duplicates 6%, video dependence +7%, protocol issues 8% of errors, coverage pool 78, Video-Index 4

S-EMBER

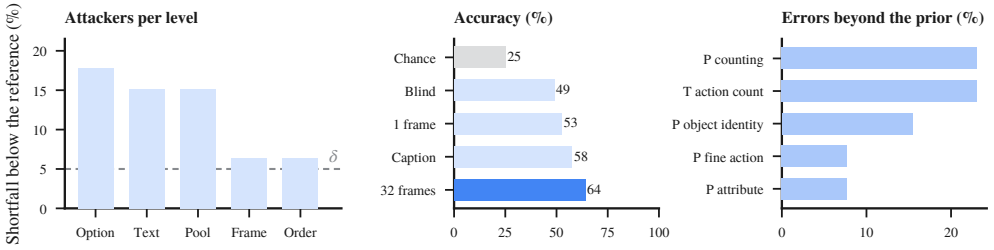
released 2026, chance 20%, reference 37% at 32 frames, survives every level



Effective size 43 of 200, near-duplicates 12%, video dependence +8%, protocol issues 0% of errors, coverage pool 505, Video-Index 109

ScaleLong

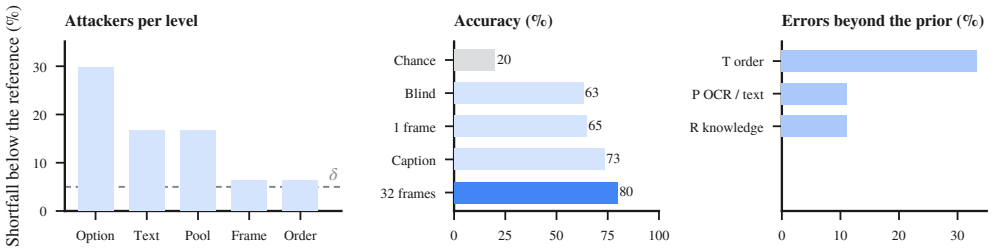
released 2025, chance 25%, reference 64% at 32 frames, survives every level



Effective size 53 of 200, near-duplicates 28%, video dependence +15%, protocol issues 10% of errors, coverage pool 161, Video-Index 4

SLVMBench

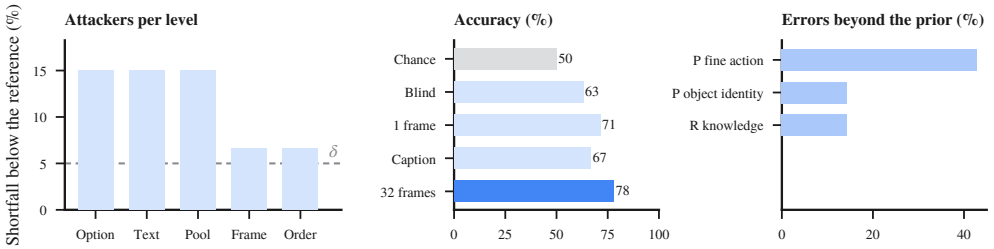
released 2026, chance 20%, reference 80% at 32 frames, survives every level



Effective size 40 of 200, near-duplicates 51%, video dependence +17%, protocol issues 7% of errors, coverage pool 257, Video-Index 21

SmartHome-Bench

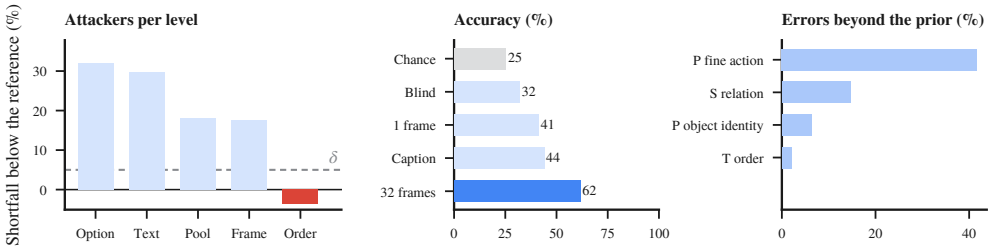
released 2025, chance 50%, reference 78% at 32 frames, survives every level



Effective size 4 of 200, near-duplicates 99%, video dependence +15%, protocol issues 29% of errors, coverage pool 18, Video-Index 1

TaskMeAnything-v1

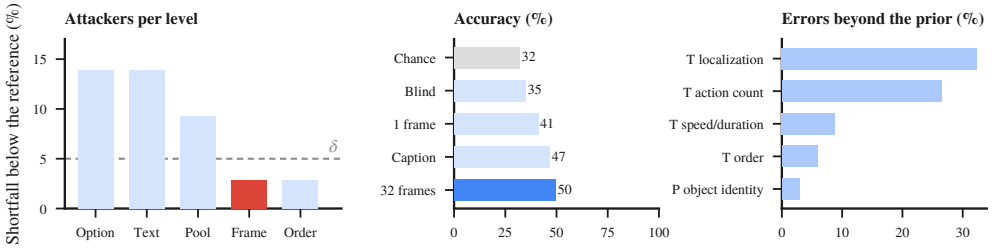
released 2024, chance 25%, reference 62% at 32 frames, breaks at the order level



Effective size 16 of 200, near-duplicates 84%, video dependence +30%, protocol issues 0% of errors, coverage pool 51, Video-Index 1

TOC-Bench

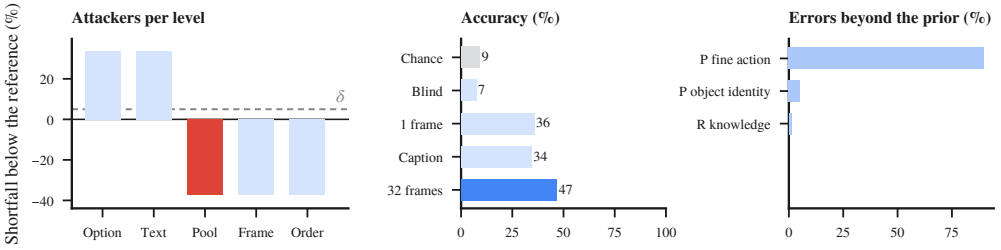
released 2026, chance 32%, reference 50% at 32 frames, breaks at the frame level



Effective size 33 of 200, near-duplicates 42%, video dependence +15%, protocol issues 1% of errors, coverage pool 97, Video-Index 38

UCF101-AD

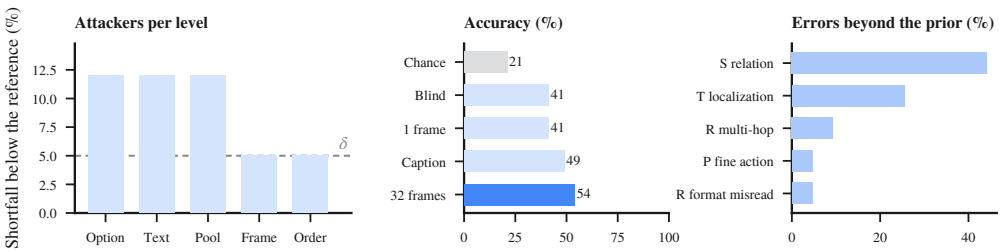
released 2026, chance 9%, reference 47% at 32 frames, breaks at the pool level



Effective size 13 of 200, near-duplicates 92%, video dependence +39%, protocol issues 21% of errors, coverage pool 5, Video-Index 0

UrbanVideo-Bench

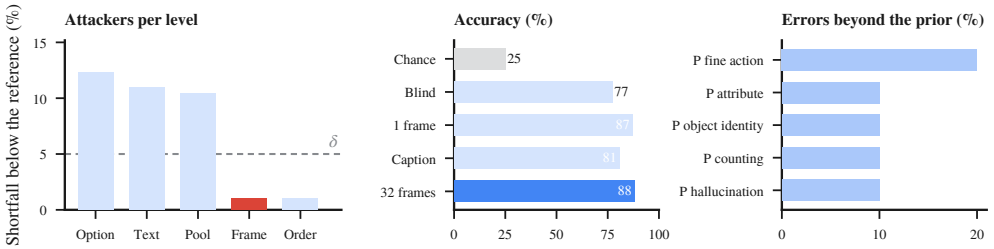
released 2025, chance 21%, reference 54% at 32 frames, survives every level



Effective size 13 of 200, near-duplicates 94%, video dependence +13%, protocol issues 1% of errors, coverage pool 85, Video-Index 34

VANE-Bench

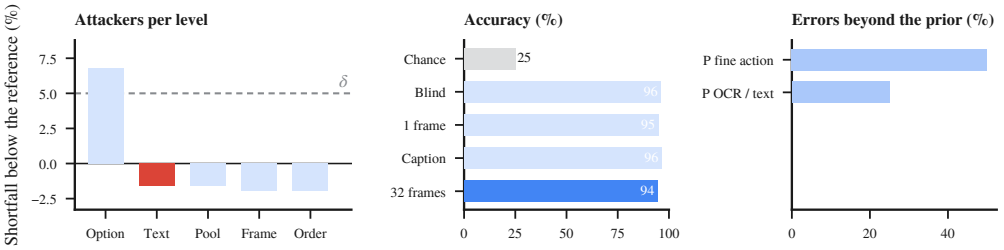
released 2025, chance 25%, reference 88% at 32 frames, breaks at the frame level



Effective size 31 of 200, near-duplicates 72%, video dependence +11%, protocol issues 0% of errors, coverage pool 39, Video-Index 1

VAU-Bench

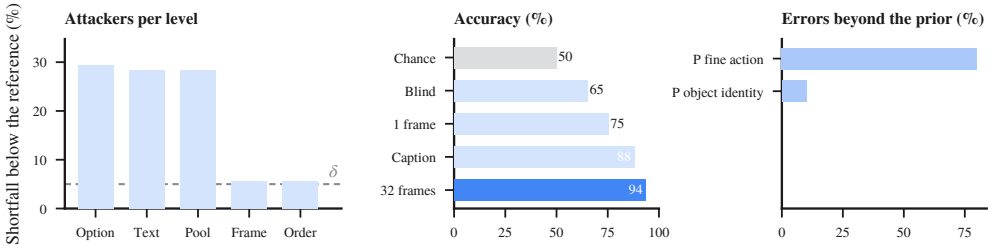
released 2025, chance 25%, reference 94% at 32 frames, breaks at the text level



Effective size 23 of 200, near-duplicates 70%, video dependence -2%, protocol issues 0% of errors, coverage pool 2, Video-Index 0

VELOCITI

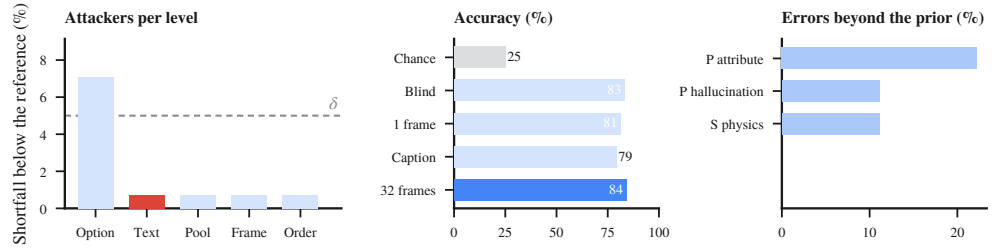
released 2025, chance 50%, reference 94% at 32 frames, survives every level



Effective size 4 of 200, near-duplicates 99%, video dependence +28%, protocol issues 0% of errors, coverage pool 16, Video-Index 11

VF-Eval

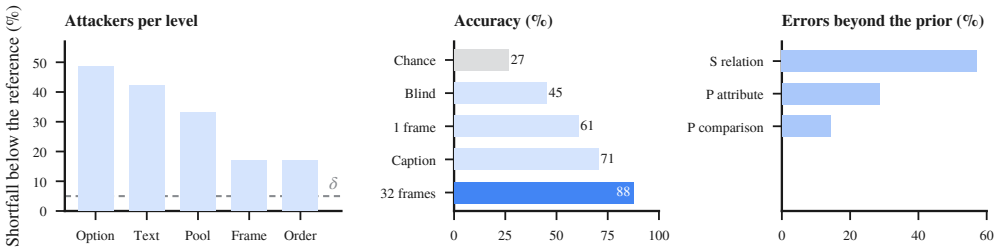
released 2025, chance 25%, reference 84% at 32 frames, breaks at the text level



Effective size 37 of 200, near-duplicates 30%, video dependence +1%, protocol issues 0% of errors, coverage pool 139, Video-Index 18

VGenST-Bench

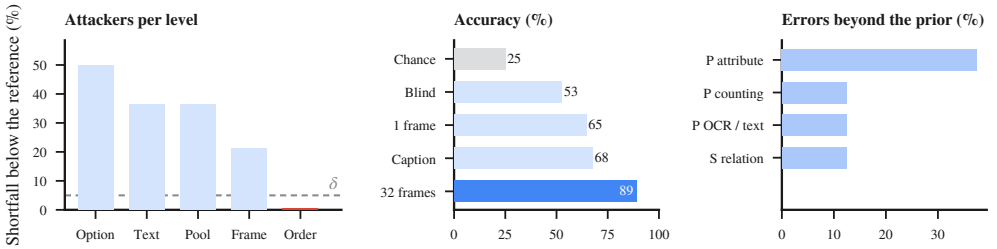
released 2026, chance 27%, reference 88% at 32 frames, survives every level



Effective size 50 of 200, near-duplicates 28%, video dependence +42%, protocol issues 0% of errors, coverage pool 216, Video-Index 6

Video-LevelGauge

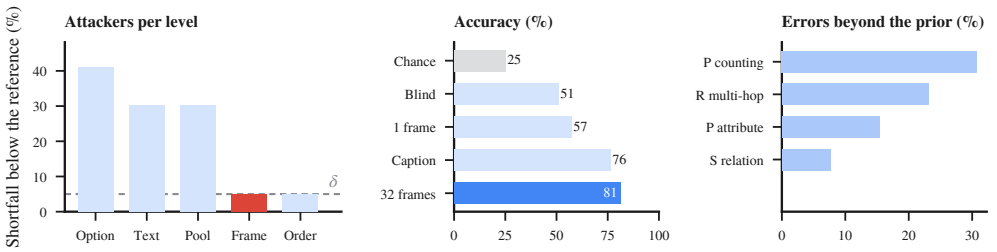
released 2026, chance 25%, reference 89% at 32 frames, breaks at the order level



Effective size 68 of 200, near-duplicates 22%, video dependence +36%, protocol issues 0% of errors, coverage pool 74, Video-Index 2

VideoAds

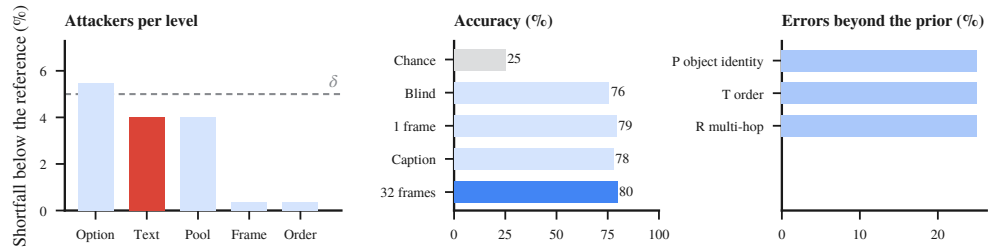
released 2025, chance 25%, reference 81% at 32 frames, breaks at the frame level



Effective size 40 of 200, near-duplicates 27%, video dependence +30%, protocol issues 23% of errors, coverage pool 101, Video-Index 31

VideoEspresso

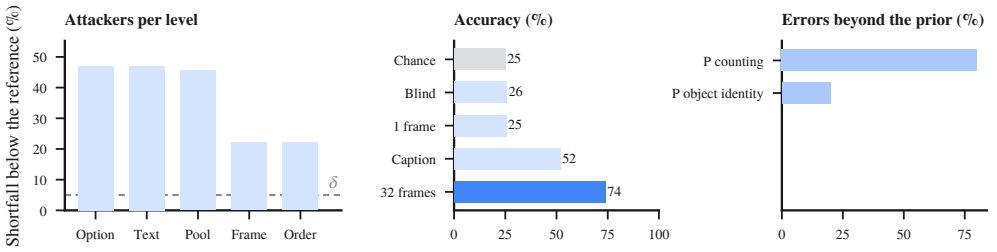
released 2025, chance 25%, reference 80% at 32 frames, breaks at the text level



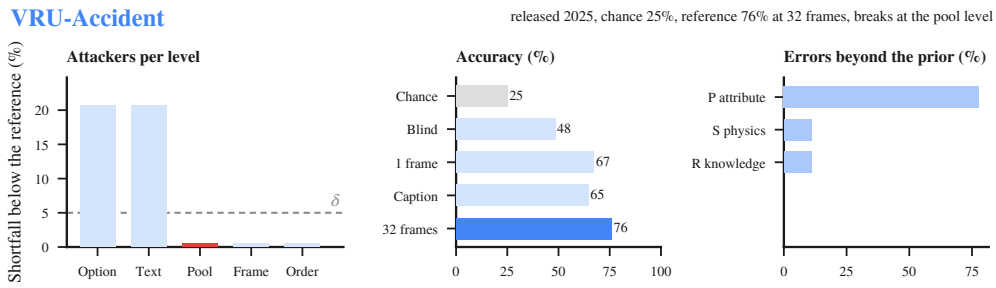
Effective size 52 of 200, near-duplicates 23%, video dependence +4%, protocol issues 12% of errors, coverage pool 116, Video-Index 7

VNBench

released 2025, chance 25%, reference 74% at 32 frames, survives every level



Effective size 25 of 200, near-duplicates 58%, video dependence +48%, protocol issues 4% of errors, coverage pool 37, Video-Index 0



Effective size 5 of 200, near-duplicates 100%, video dependence +28%, protocol issues 9% of errors, coverage pool 30, Video-Index 11